PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G01N 33/68

A1

(11) International Publication Number: **WO 99/06839**

(43) International Publication Date: 11 February 1999 (11.02.99)

(21) International Application Number: PCT/US98/15943

(22) International Filing Date: 30 July 1998 (30.07.98)

(30) Priority Data:
08/904,842 1 August 1997 (01.08.97) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US 08/904,842 (CON)
Filed on 1 August 1997 (01.08.97)

(71) Applicant *(for all designated States except US)*: NOVALON PHARMACEUTICAL CORPORATION [US/US]; Suite 560, 4222 Emperor Boulevard, Durham, NC 27703–8466 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: THORP, H., Holden [US/US]; 215 Marilyn Lane, Chapel Hill, NC 27514 (US).

(74) Agent: COOPER, Iver, P.; Browdy and Neimark, P.L.L.C., Suite 300, 419 Seventh Street N.W., Washington, DC 20004 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD OF IDENTIFYING AND DEVELOPING DRUG LEADS

(57) Abstract

A protein of interest is characterized by reactivity and/or aptamer descriptors as similar to a database protein, whose activity is mediated by a known drug. The search for drugs which mediate the protein of interest emphasizes compounds similar in structure to those which mediate the activity of the higher–scoring database proteins.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

METHOD OF IDENTIFYING AND DEVELOPING DRUG LEADS

BACKGROUND OF THE INVENTION

Field of the Invention

This invention relates to an improvement in the art of using combinatorial chemistry to develop drug leads.

Description of the Background Art

Drug Discovery. The human genomics effort could yield gene sequences that code for as many as 70,000 proteins, each a potential drug target; microbial genomics will increase this number further. Unfortunately, since genomic studies identify genes, but not the biological activity of the corresponding proteins, it is likely that many of the genes will prove to encode proteins whose activation or inactivation has no effect on disease progression. (Gold, et al., J. Nature Biotech., 15:297, 1997). There is therefore a need for a method of determining which proteins are most likely to be productive targets for pharmacological intervention.

Even if one knew in advance the perhaps 10,000 proteins which could be considered interesting targets, there remains the problem of efficiently screening hundreds of thousands of possible drugs for a useful activity against these 10,000 targets.

Historically, acquiring chemical compound libraries has been a barrier to the entry of smaller firms into the drug discovery arena. Due to the large quantity of chemical required for testing on whole animals and even on cells in culture, it was a given that whenever a compound was synthesized it should be done in fairly large quantity. Thus, there was a synthesis and purification throughput of less than 50 compounds per chemist per year. Large companies maintained their immensely valuable collections as trade barriers. However, with the downsizing of targets to the molecular level and the automation of screens, the quantity of a given compound necessary for an assay has been reduced to very small amounts. These changes have opened the door for the utilization of so-

called combinatorial chemistry libraries in lieu of the traditional chemical libraries. Combinatorial chemistry permits the rapid and relatively inexpensive synthesis of large numbers of compounds in the small quantities suitable for automated assays directed at molecular targets. Numerous small companies and academic laboratories have successfully engineered combinatorial chemical libraries with a significant range of diversity (reviewed in Doyle, 1995, Gordon *et al*, 1994a, Gordon *et al*, 1994b).

Combinatorial Libraries. In a combinatorial library, chemical building blocks are randomly combined into a large number (as high as 10E15) of different compounds, which are then simultaneously screened for binding (or other) activity against one or more targets.

Libraries of thousands, even millions, of random oligopeptides have been prepared by chemical synthesis (Houghten *et al.*, Nature, 354:84-6(1991)), or gene expression (Marks *et al.*, J Mol Biol, 222:581-97(1991)), displayed on chromatographic supports (Lam *et al.*, Nature, 354:82-4(1991)), inside bacterial cells (Colas *et al.*, Nature, 380:548-550(1996)), on bacterial pili (Lu, Bio/Technology, 13:366-372(1990)), or phage (Smith, Science, 228:1315-7(1985)), and screened for binding to a variety of targets including antibodies (Valadon *et al.*, J Mol Biol, 261:11-22(1996)), cellular proteins (Schmitz *et al.*, J Mol Biol, 260:664-677(1996)), viral proteins (Hong and Boulanger, Embo J, 14:4714-4727(1995)), bacterial proteins (Jacobsson and Frykberg, Biotechniques, 18:878-885(1995)), nucleic acids (Cheng *et al.*, Gene, 171:1-8(1996)), and plastic (Siani *et al.*, J Chem Inf Comput Sci, 34:588-593(1994)).

Libraries of proteins (Ladner, USP 4,664,989), peptoids (Simon *et al.*, Proc Natl Acad Sci U S A, 89:9367-71(1992)), nucleic acids (Ellington and Szostak, Nature, 246:818(1990)), carbohydrates, and small organic molecules (Eichler *et al.*, Med Res Rev, 15:481-96(1995)) have also been prepared or suggested for drug screening purposes.

The first combinatorial libraries were composed of

peptides or proteins, in which all or selected amino acid positions were randomized. Peptides and proteins can exhibit high and specific binding activity, and can act as catalysts. In consequence, they are of great importance in biological systems. Unfortunately, peptides *per se* have limited utility for use as therapeutic entities. They are costly to synthesize, unstable in the presence of proteases and in general do not transit cellular membranes. Other classes of compounds have better properties for drug candidates.

Nucleic acids have also been used in combinatorial libraries. Their great advantage is the ease with which a nucleic acid with appropriate binding activity can be amplified. As a result, combinatorial libraries composed of nucleic acids can be of low redundancy and hence, of high diversity. However, the resulting oligonucleotides are not suitable as drugs for several reasons. First, the oligonucleotides have high molecular weights and cannot be synthesized conveniently in large quantities. Second, because oligonucleotides are polyanions, they do not cross cell membranes. Finally, deoxy- and ribo-nucleotides are hydrolytically digested by nucleases that occur in all living systems and are therefore usually decomposed before reaching the target.

There has therefore been much interest in combinatorial libraries based on small molecules, which are more suited to pharmaceutical use, especially those which, like benzodiazepines, belong to a chemical class which has already yielded useful pharmacological agents. The techniques of combinatorial chemistry have been recognized as the most efficient means for finding small molecules that act on these targets (3). At present, small molecule combinatorial chemistry involves the synthesis of either pooled or discrete molecules that present varying arrays of functionality on a common scaffold (4). These compounds are grouped in libraries that are then screened against the target of interest either for binding or for inhibition of biological activity. Libraries containing hundreds of thousands of compounds are now being routinely synthesized; however, screening these large

libraries for binding or inhibition with all 10,000 potential targets cannot be reasonably accomplished with present screening technologies, and there are numerous experimental and computational strategies under development to reduce the number of compounds that must be screened for each target (*5-8*).

Information-intensive drug discovery. As pointed out by Paterson, et al., J. Med. chem., 39: 3049-59 (1996), medicinal chemistry advances through the dual processes of "lead discovery" and "lead optimization". In "lead discovery", the search objective is the discovery of an "activity island", a chemical class with a high frequency of active molecules. (this class may be defined mathematically as a volume within a multidimensional space defined by various molecular descriptors). In "lead optimization", the "activity island" is explored in detail. If each compound synthesized and tested can be considered as a probe of a "neighborhood" of similar compounds, in "lead discovery", it is inefficient to test compounds whose neighborhoods overlap.

Coupled to the recent advancements in genomics and molecular biology has been a revolution in information technology, which includes relational databases, computer graphics, and neural networks (*13*). These capabilities permit the construction of databases of descriptors that describe either compounds or targets in quantitative terms, and these descriptors can be related to make predictions about the structures of compounds, their biological activities, and the targets they act on (*5-8*).

Structure descriptors can be based on a variety of structural features. These approaches provide arrays of molecular descriptors that can be used to assess the similarity of molecules in a library.

Paterson et al. (1996) ranked 11 molecular diversity descriptors according to their utility in defining a neighborhood region. In order of increasing usefulness, these were random numbers =log P= MR =strain energy < connectivity indices < 2D fingerprints (whole molecule) = atom pairs = autocorrelation indices < steric CoMFA = 2D fingerprints (side chain only) = H-bonding CoMFA fields. The authors note that

any group of individual diversity descriptors can be combined into one composite descriptor by analogy to Euclidean distance, where composite distance is the square root of the weighted sum of the squares of the individual descriptors. They suggest autoscaling the vector by dividing each individual descriptor by its observed standard deviation.

Klebe and Abraham, J. Med. Chem., 36: 70-80 (1993) explored the ability of comparative molecular field analysis to predict new biologically active compounds on the basis of previously tested compounds. In essence, this method involves postulating a drug-protein alignment and then calculating the steric and electrostatic interactions between the two at regularly spaced grid points of a 3D lattice. The authors found that enthalpies, but not free enthalpies or binding constants, could be predicted by this method for molecules binding to human rhinovirus 14, thermolysin, and renin.

To facilitate the selection of structures from large compound sets for combinatorial chemical synthesis and high throughput automated bioassays, Cummins, et al., J. Chem. Inf. Comput. Sci. 36: 750-63 (1996) first calculated 109 different structural descriptors for each of over 300,000 compounds (these were taken from two databases of commercially available compounds and one of compounds proprietary to the Wellcome Foundation). Favoring the descriptors with a greater degree of independence and normality of distribution, they winnowed the list down to a final set of 61 descriptors (one physical property, the free energy of solvation, and 60 topological indices, such as the number of vertices). Together, these defined a "descriptor space" into which each compound could be placed. Factor analysis was used to reduce the dimensionality of the data; only four factors were needed to explain 90% of the variation in the data, and eight factors to explain over 95%. Only the top six eigenvalues were greater than unity.

The authors then classified compounds from two databases of compounds with medicinal activity (CMC, "Comprehensive Medicinal Chemistry", and MDDR, "MACCS-II Drug Data Report") into the same descriptor space. They suggested that compounds in the commercially available compound database which, in

descriptor space, overlap with the descriptor space of the biologically active compounds, are of interest for drug development.

Matter, J. Med. Chem., 40:1219-29 (1997) was interested in selecting an ensemble of nonredundant compounds for mass screening. The author asked, "Which physicochemical measure of similarity correlates with biological properties?" Matter evaluated a variety of molecular descriptors, including 2d fingerprints, atom-pair fingerprints, topological 2D descriptors, autocorrelation functions for atomic properties, flexible 3D fingerprints, molecular shape descriptors, and WHIM (weighted holistic invariant molecular) indices. Cluster analysis was performed on the 1283 biologically active compounds, in 55 bioactivity classes, from the IndexChemicus93 database. The ability of each descriptor to predict the biological activity of one compound based on its similarity (measured by that descriptor) to another compound was evaluated by a chi-squared statistical test. The 2D fingerprint descriptors were found to be the most useful in making predictions.

Such information is useful in combinatorial chemistry, because in *optimizing* leads, testing similar compounds is productive, while in *discovering* leads, testing similar compounds is wasteful (7). Therefore, a quantitative description of the similarities of compounds based on their structures (and by necessity a quantitative understanding of what "similar" means) can be used to direct efficient drug discovery. Indeed, quantitative structure-activity relationships show that many of these descriptors can in fact be correlated with biological activity of the compounds in the library.

An exciting recent application of this approach has been described by the National Cancer Institute for the molecular pharmacology of cancer (*13*). In this approach, there are three databases that are related.

The "activity" database (**A**) contains the activities against 60 cell lines for 60,000 compounds that have been screened at NCI. The similarity in the activity profile

against the panel of cell lines can then be calculated for any two compounds, and is generally assessed by a pairwise correlation coefficient (PCC), which is determined by an algorithm called COMPARE, which calculates the similarity of all of the compounds in the database to a user-supplied "seed" compound.

The "target" database (**T**) has been created for 100 proteins (targets) whose level of expression was determined in the same 60 cell lines. These expression levels were assessed by standard biological techniques that determine either the quantity of expressed protein (e.g., by Western blots or immunocytochemistry) or the quantity of messenger RNA (e.g., by quantitative PCR or Northern blots) for each protein in each cell line. Relation of the **A** and **T** databases then provides information on the molecular pharmacology of the compounds in **A**; inhibition of one of the heavily expressed proteins emerges as a possible mechanism for the activity of the compound.

Finally, a "structure" database (**S**) has been compiled that contains structural descriptors for a library of 460,000 compounds that includes the compounds in **A**. Similarities between the structural descriptors can be calculated for all of the compounds in **S**, so for a given active compound in **A**, unscreened, but structurally similar, compounds can be identified in **S**. These unscreened compounds have an increased likelihood of being active in the cell lines for which the screened compounds are active. The latter process therefore provides a means for "lead optimization" after a compound with a given biological activity has been identified. The NCI approach in defining the target database (**T**) is significantly different from that described here in that it relies solely on biological activity assays.

For proteins, structural descriptors cannot be directly calculated from the amino acid sequence, because the three-dimensional structure is not known and the residues that comprise the binding site are not known. We describe here a database for protein targets that will have the same predictive value as chemical library databases in predicting similarities between proteins; however, the quantitative descriptors will

be determined by *in vitro* experiments rather than from calculation.

While there are numerous computational approaches for *in vitro* typing of small molecules based on their chemical structures (*5-8*), there are no analogous experimental or theoretical methods for obtaining *in vitro* information on protein targets that can be used to relate similar proteins, outside of actually screening small molecules (*5*).

Kauvar, et al., Chemistry & Biology, 2: 107-118 (1995) "fingerprinted" over 5,000 compounds by the binding potency (concentration needed to inhibit 50% of the protein's activity) of each compound to each member of a reference panel of eight proteins. (These proteins were selected on the basis of readily assayable activity, broad cross-reactivity with small organic molecules, and low correlation between each other in binding patterns.) A screening library of 54 compounds was then selected based on the diversity in their "fingerprints" (inhibitory activity against the reference panel proteins).

This "training set" was used to evaluate the similarity of the ligand binding characteristics of a new protein to one of the reference panel proteins. By regression analysis, a computational surrogate (a weighted sum of two or more reference panel proteins) for the new protein is determined. The activity of all fingerprinted compounds to inhibit the activity of the new protein is predicted as the sum of their appropriately weighted inhibitory activities against the component reference proteins of the computational surrogate. Predictions may be improved by testing additional sets of compounds against the new protein. See also L. M. Kauvar, H. O. Villar. Method to identify binding partners. US Patent 5587293.

In one embodiment of the present method, proteins are fingerprinted on the basis of their chemical reactivity, in the presence or absence of a binding partner, rather than on the basis of their biological activity. Therefore, we do not need to identify a large number of diverse affinity molecules for each reference protein.

In another embodiment, proteins are fingerprinted on the

basis of their affinity for peptides or nucleic acids in a high-diversity library. This library provides a far greater range of conformational variation than is provided by Kauvar's training set.

Biomolecule reactivity. Chemical reactions that modify nucleotides have been extremely successful in probing the structures of complex DNA and RNA molecules (*16,17*). In these studies, a reagent that oxidizes nucleic acids by a particular reaction pathway is used to create backbone lesions in the polyanion. The sites of modification are then determined by high-resolution gel electrophoresis. These sites then indicate where the reactive functionality on the nucleotide (e.g., guanine N7 or deoxyribose C4') is exposed to the solution. Despite the success of these studies in defining complex nucleic acid structures, these concepts have not been used to define protein structures, primarily because the amide backbone is much more difficult to cleave than the phosphodiester backbone (*11,12*). In other words, most reactions that damage nucleotides cleave the phosphodiester backbone, but reactions that damage amino acid side chains generally do not cleave the amide backbone. Thus, high-resolution gel electrophoresis cannot be used to map the sites where amino acids are reactive toward a given reagent and therefore exposed to solution.

A diplatinum compound, PtPop, has been used to footprint DNA. Breiner, K. M., M. A. Daugherty, T. G. Oas and H. H. Thorp (1995). "An Anionic Diplatinum DNA Photocleavage Agent: Chemical Mechanism and Footprinting of Lambda Repressor." J. Am. Chem. Soc. **117**: 11673-11679.

Chao, "Modification of Protein Surface Hydrophobicity and Methionine Oxidation by Oxidative Systems.", Proc. Nat. Acad. Sci. USA 94: 2969-74 (1997) discovered that the surface hydrophobicity of rat liver proteins increases with animal age, and that in vitro exposure of such proteins to a metal-catalyzed oxidation system (ascorbate/Fe(II)/hydrogen peroxide) or to a peroxyl radical-generating system (AAPH:2,2'-azobis(2-amidinopropane) dihydrochloride) leads to an increase in surface hydrophobicity, protein carbonyl content, and conversion of methionine to methionyl sulfoxide. The AAPH

system also resulted in oxidation of tryptophan residues, precipitation of some proteins, and formation of dityrosine derivatives. Changes in surface hydrophobicity were detected by means of the change in protein fluorescence (490 nm) associated with binding of ANSA (8-anilino-1-naphthalene-sulfonic acid) to protein surface hydrophobic regions. Chao et al. suggested use of surface hydrophobicity measurements as a marker for protein age.

Levine et al, "Methionine Residues as Endogenous Antioxidants in Proteins.", Proc. Nat. Acad. Sci. (USA) 93: 15036-40 (1996) comments that all amino acids are subject to oxidation, although their susceptibilities vary greatly. Methionine residues, they note, are especially susceptible (with susceptibility generally correlating with their degree of surface exposure), and a repair mechanism (via endogenous methionine sulfoxide reductases) for restoring oxidative damage to methionine residues is widespread. Levine et al. suggested that the surface methionine residues act as an antioxidant defense, scavenging oxidants before they can attack residues critical to structure or function. Levine et al. oxidized glutamine synthetase by a metal-catalyzed system, and found that a significant number of methionine residues could be oxidized without an increase in proteolytic susceptibility. Levine et al. suggested engineering proteins to increase the number of surface methionines (for longer half-life).

No one has previously suggested that the chemical reactivity of a protein may be used as a descriptor for that protein, or that the chemical reactivity of a ligand-protein complex may be used as a descriptor for that ligand and its protein binding site.

All references, including any patents or patent applications, cited in this specification are hereby incorporated by refernce. No admission is made that any reference constitutes prior art. The discussion of the references states what their authors assert and applicants reserve the right to challenge the accuracy and pertinency of the cited document.

SUMMARY OF THE INVENTION

In one embodiment of the present invention, a prospective "query" protein, usually of unknown structure, is characterized by a "reactivity descriptor", by which it is related to previously studied proteins (called "library" or "reference", or "database" proteins) for which both a "reactivity descriptor" and one or more drug leads are known. A combinatorial chemical library enriched in or even limited to chemical compounds similar to the drug leads previously identified for the related proteins in the database is then synthesized and screened for binding or other activity against the "query protein". This invention thereby reduces the number of chemical compounds that must be screened against an unknown protein target, and/or increases the likelihood of "hits".

The reactivity descriptors here contemplated relate to the reactivity of the target protein (the term "target protein" refers to both "query" and "reference" proteins, both being described by their chemical reactivity), especially in both a ligand bound and in a free state, with one or more chemical reagents. For a given reagent, the difference in the two reactivities is characteristic of the part of the protein which is occluded or otherwise shielded as a result of the ligand binding. The protected portion of the protein will include the actual ligand binding site. One may also compare the reactivity in the bound state with that of the target protein in the unfolded state, and that of the protein in a free but folded state to that which it enjoys in the unfolded state. These comparisons provide further information about the structure of the protein.

The ligands used to define the binding site of the protein may be the natural ligands therefor, if known, or they may be "surrogate ligands". A surrogate ligand is a molecule which binds the target protein, and has the potential of binding it at a site at which the target protein is bound by one of its natural ligands. Surrogate ligands may be obtained by synthesizing and screening combinatorial libraries. Since the surrogate ligand will be used only for gathering structural information, and not as a drug per se, the library may be

chosen from the point of view of obtaining the most structural diversity for the least synthetic effort, ignoring the suitability of the library members as drugs. For this reason, a preferred surrogate ligand library is a peptide ("BioKey") or oligonucleotide library.

In another embodiment of the invention, a query protein is related to reference proteins by its ability to bind similar surrogate ligands. A combinatorial oligomeric library of surrogate ligands, typically peptides or nucleic acids, is screened, and the oligomers which bind the target protein (and thus are called "aptamers") are characterized to yield "aptamer" or descriptors. The aptamer descriptors (sequences and, possibly, additional information such as contact points and secondary structure) identified for the query protein are compared to those identified for the database proteins, and the drug leads previously identified for the database proteins characterized by the most similar surrogate ligands are favored.

In a preferred embodiment, the aptamer are nucleic acids, and the bases involved in protein binding are determined by footprinting techniques. By taking into account the predicted secondary structure of the nucleic acids, the epitope of the nucleic acid may be characterized as a sequence whose elements are unpaired G, A, T, or C, or any of the sixteen possible pairings (matched or mismatched) of those four bases. This sequence, for each surrogate ligand binding a reference protein, may be compared to that of the epitope of each aptamer binding the query protein.

It is especially preferred that the characterization of the proteins by reactivity descriptors be combined with its characterization by aptamer descriptors (surrogate ligand binding).

If desired, the work involved may be reduced by first screening a potential surrogate ligand library to obtain "aptamer descriptors" for the target protein, and then using the bound molecules (aptamers) from that library to modulate the chemical reactivity of the protein and thereby help characterize its binding site(s) by means of chemical

reactivity descriptors. However, our preference is to use peptides to alter reactivity and nucleic acids to generate aptamer descriptors.

On the basis of reactivity and/or aptamer-based descriptors, the similarity of the query protein to each of the reference proteins is determined. For each reference protein, one or more drug leads are known, so these drug leads may be rated or ranked, as drug leads for modulators of the query protein, based on the similarity of their reference protein to the query protein, and, optionally, their own drug characteristics (e.g., potency, half-life, side effects).

A combinatorial library is then synthesized which is enriched for members which are structurally similar to the aforementioned drug leads. Structurally similar members may be identified in a formal manner by use of the chemical structure descriptors available in the art, or more informally through a chemist's expert judgment of structural similarity.

The chemist may also develop the drug leads without resorting to a combinatorial library, i.e., by synthesizing lead analogues on an individual, noncombinatorial basis.

The lead analogues, whether in the form of a combinatorial library or discretely prepared compounds, are then screened for the ability to modulate the target protein's activity, <u>in vitro</u> or <u>in vivo</u>. Successful analogues are added to the database as leads associated with the target protein, which now becomes a reference protein.

Thus, the initial lead discovery occurs through querying a database of reference proteins and their associated descriptors and drugs. The combinatorial library is screened primarily for purpose of optimization of these leads, although it is likely to be sufficiently different in structure from the lead so that there will be some secondary lead discovery as well.

The appended claims are hereby incorporated by reference as a description of the preferred embodiments.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1. Cartoon showing the reactivity of the transition-metal reagents toward nucleotides. $Ru(typ)(bpy)O^{2+}$ (RuO) reacts by hydrogen abstraction at 1′ and oxygen transfer at guanine C8. $Pt_2(pop)_4^{4-}$ reacts by hydrogen abstraction at 4′ and 5′ by outer-sphere electron transfer at guanine. $Ru(bpy)_3^{3+}$ reacts only by outer-sphere electron transfer. The reactivity of the complexes towards amino acids should therefore be different for all three reagents as well.

Figure 2. Amino acids likely to be reactive by hydrogen abstraction. Probable sites of C-H activation are shown in boxes.

Figure 3. Scheme showing the solvent accessibility of reactive amino acid residues (*) as a function of folding and binding of the BioKey peptide.

Figure 4. Scheme showing the assay used to determine the relative rates for modification of a protein by a reagent. The presence of the protein decreases the yield of Form II DNA in a manner related to the rate constant for oxidation of the protein by the reagent.

Figure 5. Fig. 5A shows the secodary structure of an RNA sequence, with protein contact sites marked with an arrow. Fig. 5B is a two-dimensional grid representation of the same information.

Figure 6. Flow Chart of Preferred Method.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS OF THE INVENTION

The present invention is directed to a method for the more efficient identification of small organic molecules, preferably molecules having a molecular weight of less than 500 daltons, which are pharmaceutically acceptable and which are potent modulators of the biological activity of a protein.

Receptor-Mediated Pharmacological Activity

Many pharmacologically active substances elicit a physiological response by interacting with a specialized portion, known as a receptor, of the target cell. The substances which are able to elicit that response, by specific interaction with a receptor site, are known as agonists. Typically, increasing the concentration of the agonist at the receptor site leads to an increasingly larger response, until a maximum response is achieved. A substance able to elicit the maximum response is known as a full agonist, and one which elicits only, at most, a lesser (but discernible) response is a partial agonist.

A pharmacological antagonist is a compound which interacts with the receptor but without eliciting a response. By doing so, it inhibits the receptor from responding to agonists. A competitive antagonist is one whose effect can be overcome by increasing the agonist concentration; a noncompetitive antagonist is one whose action is unaffected by agonist concentration.

Ligands are substances which bind to receptors, and thereby encompass both agonists and pharmacological antagonists. Ligands which activate (agonize) or inhibit (antagonize) the receptor are here termed modulators.

The clinical concept of drug antagonism is broader than the pharmacological concept, including phenomena that do not involve direct inhibition of agonist:receptor binding. A "physiological" antagonist could be a substance which directly or indirectly inhibits the production or release of the natural agonist, or directly or indirectly facilitates its elimination

from the receptor site. A physiological antagonist of one receptor may be a pharmacological agonist of another receptor, such as one which activates an enzyme which degrades the natural ligand of the first receptor.

If a disease state is the result of inappropriate activation of a receptor, the disease may be prevented or treated by means of a physiological or pharmacological antagonist. Other disease states may arise through inadequate activation of a receptor, in which case the disease may be prevented by means of a suitable agonist.

An important class of receptors are proteins embedded in the phospholipid bilayer of cell membranes. The binding of an agonist to the receptor (typically at an extracellular binding site) can cause an allosteric change at an intracellular site, altering the receptor's interaction with other biomolecules. The physiological response is initiated by the interaction with this "second messenger" (the agonist is the "first messenger") or "effector" molecule.

The peptides and nucleic acids used in the present invention can act as agonists (binding to the receptor and causing its activation), as antagonists (binding to the receptor without activating it, and blocking its activation by agonists), or as pharmacologically neutral species (binding to the receptor without either activating or blocking it).

Enzymes are special types of receptors. Receptors interact with agonists to form complexes which elicit a biological response. Ordinary receptors then release the agonist intact. With enzymes, the agonists are enzyme substrates, and the enzymes catalyze a chemical modification of the substrate. Enzymes are not necessarily integral membrane proteins; they may be secreted, or intracellular, proteins. Often, enzymes are activated by the action of a receptor's second messenger, or, more indirectly, by the product of an "upstream" enzymatic reaction.

Thus, drugs may also be useful because of their interaction with enzymes. The drug may serve as a substrate for the enzyme, as a coenzyme, or as an enzyme inhibitor. (An irreversible inhibitor is an "inactivator".) Drugs may also

cause, directly or indirectly, the conversion of a proenzyme into an enzymes. Many disease states are associated with inappropriately low or high activity of particular enzymes.

The present invention may be used to identify both agonists and antagonists of receptors. It is not unusual for a relatively small structural change to convert an agonist into a pharmacological antagonist, or vice versa. Therefore, even if the drugs known to interact with a reference protein are all agonists, the drugs in question may serve as leads to the identification of both agonists and antagonists of the reference protein and of related proteins. Similarly, known antagonists may serve as drug leads, not only to additional antagonists, but to agonists as well.

The nucleic acid aptamers and BioKey peptides used in developing descriptors are selected only for their ability to bind the target protein; it is not required that they activate the protein, or inhibits its activation. Some will bind at sites at which agonism/antagonism can occur, others will not. Their purpose is to help characterize the target protein surface, not to serve as drug leads themselves (although the practitioner is free to test the nucleic acids and peptides for agonist/antagonist activity and to use the active ones as leads in the design of active analogues which are more suitable than nucleic acids and peptides per se as drugs).

Protein Binding and Biological Activity

Many of the biological activities of the proteins are attributable to their ability to bind specifically to one or more binding partners (ligands), which may themselves be proteins, or other biomolecules.

When the binding partner of a protein is known, it is relatively straightforward to study how the interaction of the binding protein and its binding partner affects biological activity. Moreover, one may screen compounds for the ability of the compound to competitively inhibit the formation of the complex, or to dissociate an already formed complex. Such inhibitors are likely to affect the biological activity of the protein, at least if they can be delivered in vivo to the site

of the interaction.

If the binding protein is a receptor, and the binding partner an effector of the biological activity, then the inhibitor will antagonize the biological activity. If the binding partner is one which, through binding, blocks a biological activity, then an inhibitor of that interaction will, in effect, be an agonist of the biological activity in question.

The residues whose functional groups participate in the ligand-binding interactions together form the ligand binding site, or paratope, of the protein. Similarly, the functional groups of the ligand which participate in these interactions together form the epitope of the ligand.

In the case of a protein, the binding sites are typically relatively small surface patches. The binding characteristics of the protein may often be altered by local modifications at these sites, without denaturing the protein.

While it is possible for a chemical reaction to occur between a functional group on a protein and one on a ligand, resulting in a covalent bond, protein-ligand binding normally occurs as a result of the aggregate effects of several noncovalent interactions. Electrostatic interactions include salt bridges, hydrogen bonds, and van der Waals forces.

What is called the hydrophobic interaction is actually the absence of hydrogen bonding between nonpolar groups and water, rather than a favorable interaction between the nonpolar groups themselves. Hydrophobic interactions are important in stabilizing the conformation of a protein and thus indirectly affect ligand binding, although hydrophobic residues are usually buried and thus not part of the binding site.

Peptides have been found to bind proteins at the same sites as those by which the proteins interact with other proteins, macromolecules and biologically significant substances e.g. nucleic acids, lipids and enzyme substrates.

Potency

The potency of an antagonist of a protein may be expressed

as an IC50, the concentration of the antagonist which causes a 50% inhibition of a protein's binding or biological activity in an in vitro or in vivo assay system. A pharmaceutically effective dosage of an antagonist depends on both the IC50 of the antagonist, and the effective concentrations of the protein and its clinically significant binding partner(s).

Potencies may be categorized as follows:

Category	IC50
Very Weak	>1 μ moles
Weak	100 n moles to 1 μ mole
Moderate	10 n moles to 100 n moles
Strong	1 p mole to 10 n moles
Very Strong	<1 p mole

Preferably, the antagonists identified by the present invention are in one of the four higher categories identified above, and are in any event more potent than any antagonist known for the protein in question at the time of filing of this application.

In a similar manner, the potency of an agonist may be quantified as the dosage resulting in 50% of its maximal effect on a receptor.

Drug Leads

The term "drug lead", as used herein, refers to a compound which is a member of a structural class which is generally suitable, in terms of physical characteristics (e.g., solubility), as a source of drugs, and which has at least some useful pharmacological activity, and which therefore could serve effectively as a starting point for the design of analogues and derivatives which are useful as drugs. The "drug lead" may be a useful drug in its own right, or it may be a compound which is deficient as a drug because of inadequate potency or undesirable side effects. In the latter case, analogues and derivatives are sought which overcome these deficiencies. In the former case, one seeks to improve the already useful drug.

Such analogues and derivatives may be identified by rational drug design, or by screening of combinatorial or

noncombinatorial libraries of analogues and derivatives.

Preferably, a drug lead is a compound with a molecular weight of less than 1,000, more preferably, less than 750, still more preferably, less than 600, most preferably, less than 500. Preferably, it has a computed log octanol-water partition coefficient in the range of -4 to +14, more preferably, -2 to +7.5.

Target Proteins

The target protein (a query or reference protein) may be a naturally occurring protein, or a subunit or domain thereof, from any natural source, including a virus, a microorganism (including bacterial, fungi, algae, and protozoa), an invertebrate (including insects and worms), or the normal or cancerous cells of a vertebrate (especially a mammal, bird or fish and, among mammals, particularly humans, apes, monkeys, cows, pigs, goats, llamas, sheep, rats, mice, rabbits, guinea pigs, cats and dogs). Alternatively, the target protein may be a mutant of a natural protein. Mutations may be introduced to facilitate the labeling or immobilization of the target protein, or to alter its biological activity (An inhibitor of a mutant protein may be useful to selectively inhibit an undesired activity of the mutant protein and leave other activities substantially intact).

The target protein may be, inter alia, a glyco-, lipo-, phospho-, or metalloprotein. It may be a nuclear, cytoplasmic, membrane, or secreted protein. It may, but need not, be an enzyme. The known binding partners (if any) of the target protein may be, inter alia, other proteins, oligo- or polypeptides, nucleic acids, carbohydrates, lipids, or small organic or inorganic molecules or ions. The biological activity or function of the target protein may be, but is not limited to, being a

kinase
- protein kinase
  - tyrosine kinase
  - Threonine kinase
  - Serine Kinase
- nucleotide kinase
- polynucleotide kinase

Phosphatase
- Protein phosphatase
- nucleotide phosphatase
- acid phosphatase
- alkaline phosphatase
- pyrophosphatase

deaminase

protease
- endoprotease
- exoprotease
- metalloprotease
- serine endopeptidase
- cysteine endopeptidase

nuclease
- Deoxyribonuclease
- ribonuclease
- endonulcease
- exonuclease

polymerase
- DNA Dependent RNA polymerase
- DNA Dependent DNA polymerase
- telomerase
- primase

Helicase

Dehydrogenase

transferase
- peptidyl transferase
- transaminase
- glycosyltransferase
- ribosyltransferase
- acetyltransferase

Hydrolase
- urease

carboxylase

isomerase
- dismutase
- rotase
- topoisomerase

glycosidase
- endoglycosidase
- exoglycosidase

deaminase

lipase

esterase

sulfatase

cellulase

lyase

reductase

synthetase

Ion Channel

DNA Binding

RNA Binding

Ligase
- RNA ligase
- DNA ligase

Adaptor or scaffolding protein

Structural protein
- fibrin(ogen)
- collagen
- elastin
- talin

Tumor Suppressor

adhesion molecule

oxygenase

oxidase
- peroxidase

chaperonin

Transporter
- electron transporter
- protein transporter
- peptide transporter
- hormone transporter
  - serotonin
  - DOPA
- nucleic acid transporter

signal transduction

neurotransmitter

structural component

of viruses
of cells
of organs
of organisms

information carrier/storage

antigen recognition protein
MHC I complex
MHC II complex

receptor
TNfα Receptor
TNFβ Receptor
β-Adrenergic Receptor
α-Adrenergic Receptor
IL-8 Receptor
IL-3 Receptor
CSF Receptor
Erythropoeitin Receptor
FAS Ligand Receptor
T-cell Receptors
B-Cell Antigen Receptor
F episilon Receptor
Growth Hormone Receptor
Nuclear Receptors
Glucocorticoid
Estrogen
Testosterone

The binding protein may have more than one paratope and they may be the same or different. Different paratopes may interact with epitopes of different binding partners. An individual paratope may be specific to a particular binding partner, or it may interact with several different binding partners. A protein can bind a particular binding partner through several different binding sites. The binding sites may be continuous or discontinuous (vis-a-vis the primary sequence of the protein).

Thus, the target (query or reference) protein, may be any protein of interest. As descriptors and drug compounds are determined for proteins, the information is added to the database.

For the purpose of validating the chemical reactivity descriptor concept of the present invention, a particularly preferred initial target protein is glutathione-S-transferase (GST), which is chosen because it has been crystallographically characterized with and without a large number of bound

inhibitors (*37*), the level of expression has been measured in NCI's 60 cell lines (*13*), the activities of many known inhibitors are available (*38*), and macroscopic quantities of peptides that bind to the active site are preparable.

Other preferred proteins for early incorporation into the database are ones that are clinically relevant, have known small-molecule inhibitors, and whose expression has been assessed by biological methods. Proteins for which peptide ligands exist and for which expression data are available at NCI include ras (*39*), src (*40*), and p53 (*41*), and other promising proteins for which both kinds of data are likely to be available in the future are the UL44 protein from cytomegalovirus, and hMDM2 protein that binds to p53 (*41*). Crystallographically characterized targets are of particular interest and utility in the early stages.

A list of agonists, antagonists, radioligands and effectors for many different receptors appears in Appendix I of King, Medicinal Chemistry: Principles and Practice, pp. 290-294 (Royal Soc'y Chem. 1994). Appendix II lists blockers for various ion channels (which are another special type of receptor). The proteins set forth in these appendices are good candidates for inclusion in the reference protein database.

It will be appreciated that once a ligand is identified for a former query protein, it becomes a reference protein.

Combinatorial Libraries

The term "library" generally refers to a collection of chemical or biological entities which are related in origin, structure, and/or function, and which can be screened simultaneously for a property of interest.

The term "combinatorial library" refers to a library in which the individual members are either systematic or random combinations of a limited set of basic elements, the properties of each member being dependent on the choice and location of the elements incorporated into it. Typically, the members of the library are at least capable of being screened simultaneously. Randomization may be complete or partial; some positions may be randomized and others predetermined, and at

random positions, the choices may be limited in a predetermined manner. The members of a combinatorial library may be oligomers or polymers of some kind, in which the variation occurs through the choice of monomeric building block at one or more positions of the oligomer or polymer, and possibly in terms of the connecting linkage, or the length of the oligomer or polymer, too. Or the members may be nonoligomeric molecules with a standard core structure, like the 1,4-benzodiazepine structure, with the variation being introduced by the choice of substituents at particular variable sites on the core structure. Or the members may be nonoligomeric molecules assembled like a jigsaw puzzle, but wherein each piece has both one or more variable moieties (contributing to library diversity) and one or more constant moieties (providing the functionalities for coupling the piece in question to other pieces).

The ability of one or more members of such a library to recognize a target molecule is termed "Combinatorial Recognition". In a "simple combinatorial library", all of the members belong to the same class of compounds (e.g., peptides) and can be synthesized simultaneously. A "composite combinatorial library" is a mixture of two or more simple libraries, e.g., DNAs and peptides, or benzodiazepine and carbamates. The number of component simple libraries in a composite library will, of course, normally be smaller than the average number of members in each simple library, as otherwise the advantage of a library over individual synthesis is small.

Oligonucleotide Libraries

An oligonucleotide library is a combinatorial library, at least some of whose members are single-stranded oligonucleotides having three or more nucleotides connected by phosphodiester or analogous bonds. The oligonucleotides may be linear, cyclic or branched, and may include non-nucleic acid moieties. The nucleotides are not limited to the nucleotides normally found in DNA or RNA. For examples of nucleotides modified to increase nuclease resistance and chemical stability of aptamers, see Chart 1 in Osborne and Ellington, Chem. Rev.,

97: 349-70 (1997). For screening of RNA, see Ellington and Szostak, Nature, 346: 818-22 (1990).

There is no formal minimum or maximum size for these oligonucleotides. However, the number of conformations which an oligonucleotide can assume increases exponentially with its length in bases. Hence, a longer oligonucleotide is more likely to be able to fold to adapt itself to a protein surface. On the other hand, while very long molecules can be synthesized and screened, unless they provide a much superior affinity to that of shorter molecules, they are not likely to be found in the selected population, for the reasons explained by Osborne and Ellington (1997). Hence, the libraries of the present invention are preferably composed of oligonucleotides having a length of 3 to 100 bases, more preferably 15 to 35 bases. The oligonucleotides in a given library may be of the same or of different lengths.

Oligonucleotide libraries have the advantage that libraries of very high diversity (e.g., 10^{15}) are feasible, and binding molecules are readily amplified in vitro by polymerase chain reaction (PCR). Moreover, nucleic acid molecules can have very high specificity and affinity to targets.

In a preferred embodiment, this invention prepares and screens oligonucleotide libraries by the SELEX method, as described in King and Famulok, Molec. Biol. Repts., 20: 97-107 (1994); L. Gold, C. Tuerk. Methods of producing nucleic acid ligands, US#5595877; Oliphant et al. Gene 44:177 (1986).

The term "aptamer" is conferred on those oligonucleotides which bind the target protein. Such aptamers may be used to characterize the target protein, both directly (through identification of the aptamer and the points of contact between the aptamer and the protein) and indirectly (by use of the aptamer as a ligand to modify the chemical reactivity of the protein).

Peptide Library

A peptide library is a combinatorial library, at least some of whose members are peptides having three or more amino acids connected via peptide bonds. The peptides may be linear,

branched, or cyclic, and may include nonpeptidyl moieties. The amino acids are not limited to the naturally occurring amino acids.

A biased peptide library is one in which one or more (but not all) residues of the peptides are constant residues. The individual members are referred to as peptide ligands (PL). In one embodiment, an internal residue is constant, so that the peptide sequence may be written as

$$(X_{aa})_m\text{-}AA_1\text{-}(X_{aa})_n$$

Where Xaa is either any naturally occurring amino acid, or any amino acid except cysteine, m and n are chosen independently from the range of 2 to 20, the Xaa may be the same or different, and AA_1 is the same naturally occurring amino acid for all peptides in the library but may be any amino acid. Preferably, m and n are chosen independently from the range of 4 to 9.

Preferably, AA_1 is located at or near the center of the peptide. More specifically, it is desirable that m and n are not different by more than 2; more preferably m and n are equal. Even if the chosen AA_1 is required (or at least permissive) of the target protein (TP) binding activity, one may need particular flanking residues to assure that it is properly positioned. If AA_1 is more or less centrally located, the library presents numerous alternative choices for the flanking residues. If AA_1 is at an end, this flexibility is diminished.

The most preferred libraries are those in which AA_1 is tryptophan, proline or tyrosine. Second most preferred are those in which AA_1 is phenylalanine, histidine, arginine, aspartate, leucine or isoleucine. Third most preferred are those in which AA_1 is asparagine, serine, alanine or methionine. The least preferred choices are cysteine and glycine. These preferences are based on evaluation of the results of screening random peptide libraries for binding to many different TPs.

Ligands that bind to functional domains tend to have both constant as well as unique features. Therefore, by using "biased" peptide libraries, one can ease the burden of finding

ligands. Either "biased" or "unbiased" libraries may be screened to identify "BioKey" peptides for use in developing reactivity descriptors, and, optionally, peptide aptamer descriptors and additional drug leads.

Descriptors

A "descriptor" (also known as a parameter, character, variable, or variate) is a numerically expressed characteristic of a compound (which may be a protein, or a protein ligand), which helps to distinguish that compound from others. A descriptor value need not be absolutely specific to a compound to be useful. The characteristics may be pure structural characteristics (as in a "structural descriptor") or they may refer to the compound's interaction with other compounds, such as a binding interaction (as in an "aptamer descriptor") or a chemical reaction (as in a "reactivity descriptor"). "Paired Descriptors" are descriptors of the same property as measured in two different molecule. A "descriptor array", "list", or "set" is an array, list or set whose elements are different descriptors for the same molecule.

A plurality of comparable descriptors for two compounds may be used to calculate a similarity for the two compounds. The descriptors used in making this calculation in the present invention, for two proteins, will include (a) at least one chemical reactivity descriptor, and/or (b) at least one peptide or oligonucleotide affinity descriptor, and preferably both.

The similarity calculation may optionally consider other descriptors, such as structural descriptors of known ligands, as well.

A set of n-descriptors defines an n-dimensional descriptor space; each compound for which a descriptor set is available may be said to occupy a point in descriptor space. The dissimilarity of two compounds may be expressed as a distance between the two points which they occupy in descriptor space.

A similarity measure or coefficient quantifies the relationship between two individuals (compounds), given the values of a set of variates (descriptors) common to both. Similarity coefficients are usually defined to take values in

the range of 0 to 1.

One commonly used measure of similarity is the product moment correlation coefficient. Its correlation is unity whenever two profiles are parallel, regardless of how far apart they are in level. Two profiles may have correlation of +1 even if they are not parallel, provided that the two sets of scores are linearly related.

For binary descriptors, the simplest measure of similarity is the simple matching coefficient

$$s_{ij} = \frac{\text{number of matches}}{\text{number of comparisons}}$$

The Jaccard or Sneath coefficient modifies the simple matching coefficient by ignoring bits which in both i and j are zero, i.e., by ignoring negative matches (mutual absences). In other words, it is obtained by dividing the number of bits which are set in both descriptor bit strings, and dividing by the total number of bits set in either descriptor string. It is also called the unweighted Tanimoto coefficient.

The weighted Tanimoto coefficient for descriptors k and individuals i and j is:

$$S_{ij}=\frac{\sum_k w_k x_{ik} x_{jk}}{\sum_k w_k x_{ik}+\sum_k w_k x_{jk}-\sum_k w_k x_{ik} x_{jk}}$$

Gower has defined a general similarity coefficient which can be used for binary, qualitative, and quantitative data:

$$S_{ij}=\sum_{k=1}^{p} s_{ijk} / \sum_{k=1}^{p} w_{ijk}$$ for individuals i and j and descriptor k.

W_{ijk} is set to 1 if the comparison is valid for variable k, and to 0 otherwise. If $w_{ijk}=0$, then s_{ijk} is 0. For binary data, w_{ijk} and s_{ijk} are both 0 if the variable is negative in both individuals. The s_{ijk} is positive only if the binary variable is positive for both individuals. For qualitative data, $s_{ijk}=1$ if the individuals are the same for the kth character, and $s_{ijk}=0$ if they differ. For quantitative data, $s_{ijk}=1-|X_{ik}-X_{jk}|/R_k$ where X_{ik} is the value of descriptor k for

individual i, and R_k is the total range of variable k.

Descriptors may be quantitative or qualitative. Quantitative descriptors may be integers or real numbers. Qualitative descriptors divide the data into categories which may be, but need not be, expressible as having relative magnitudes. Binary descriptors are a special case of qualitative descriptors, in which there are just two categories, typically representing the presence or absence of a feature. Qualitative data for which the variates have several levels may be treated like binary data with each level of a variate being regarded as a single binary variable (i.e., an eight level variate expressed as eight bits). Or the levels may be numbered sequentially (i.e., an eight level variable expressed as three bits).

The reactivity descriptors are preferably quantitative in form. If the aptamer descriptors are expressed as nucleic acid sequences, with or without secondary structure and protein contact information, they are qualitative in form. If they are expressed as 2D fingerprints, they are a string of binary data.

Gower's coefficient, for qualitative data, only credits exact matches of the variate. For aptamers, it is more useful to evaluate the similarity of the sequences by a BLAST type analysis rather than to simply state whether aptamers are the same or different.

A distance measure is a similarity measure which is also a metric, i.e., satisfies the conditions (i) $d(x,y) \geq 0$; and $d(x,y)=0$ if $x=y$; (ii) $d(x,y)+d(y,x)$; and (iii) $d(x,z)+d(y,z) \geq d(x,y)$ (the metric or triangular inequality). Of course, the greater the distance, the less the similarity.

Descriptors may be weighted (or otherwise transformed) for any of several reasons, including:

(a) to reflect the perceived value of the descriptor for determining whether two proteins will be modulated by structurally similar drugs;

(b) to reflect the perceived reliability of the descriptor data;

(c) to correct for differences in scale between descriptors, so that a descriptor does not dominate

a similarity or distance calculation merely because its values are of higher magnitude or are spread over a greater range; and

(d) to correct for correlations between descriptors.

The raw descriptor values may be, but need not be, transformed prior to use in calculating distances. Typical transformations are (a) presence (1)/absence (0), (b) ln(x+1), (c) frequency in sample, (d) root, and (e) relative range, i.e., (value-min)/(max-min).

The raw descriptor values may be standardized (normalized) to have zero mean ($x'=x-\mu_x$) and/or unit variance ($x'=x/\sigma_x$), possibly both ($x'=(x-\mu_x)/\sigma_x$) or they be standardized (unitized) to fall into the range 0 to 1.

Descriptor weights may be adjusted empirically on the basis of specially designed test sets. A training set of proteins is identified. Descriptors are evaluated for each protein in the set. A training set of compounds, including are also tested against each compound in the set. These compounds are chosen so that, for any protein in the set, there is at least one compound which is an agonist or antagonist for it. A neural net, with the descriptor weights as inputs, is used to predict the activity of each compound against each protein, using the calculated protein similarities. For example, it will calculate the similarity of protein x to all other proteins, then treat the activities of the compounds against the other proteins as "knowns" and use it to predict the activity of the compounds against protein x. This is done repeatedly, with each protein taking on the role of protein x, in turn.

While this may seem similar to the approach of Kauvar et al., for us it is an optional method of deriving weights, to apply to chemical reactivity or aptamer descriptors, while, for Kauvar et al., the biological activities of the training set compounds against the query protein <u>are</u> its descriptors.

If the database usage is emphasizing lead discovery, then the training set proteins may be chosen for high diversity, e.g., insignificant sequence similarities and/or unrelated biological activities. If the plan is to use the library for

lead optimization, then the training set proteins might be members of a family of homologous proteins.

The coefficient of variation may be useful in comparing descriptors; it is the standard deviation divided by the mean. If there is no information available about the ultimate significance of a descriptor, one may give a greater weight to descriptors which have a larger CV and hence a more uniform distribution.

It must be emphasized that we do not require use of weighted descriptors, let alone of any particular method of deriving weights.

It is likely that some degree of correlation will exist among the descriptors. Standard mathematical methods, such as cluster analysis, principal components analysis, or partial least squares analysis, may be used to determine which descriptors are strongly correlated and to replace them with a new descriptor which is a weighted sum of the original correlated descriptors. One may alternatively choose (perhaps randomly) one of each pair of highly correlated descriptors and simply prune it, thereby reducing the amount of data which must be collected.

One way of correcting for correlation among the descriptors is for each descriptor m, calculate the average of its squared correlation coefficients with all descriptors n (including m=n, for which the coefficient is necessarily unity), and subtract this number from one to obtain a weight representing the fraction of the variation in descriptor m which is not explained by the "average" descriptor n. With this "average r^2" method, if we have four descriptors, and two are perfectly correlated to each other, and the descriptors are otherwise completely uncorrelated, the correlated descriptors will have weights of 0.5 each, and the other two will have weights of 1.0 each.

Distances may be calculated on the basis of any of a variety of distance measures known in the statistical arts.

The most commonly used distance measure is the Euclidean metric:

$$d_{ij}=(\sum_{k}(X_{ik}-X_{jk})^2)^{1/2}$$

It corresponds most closely to our intuitive sense of distance.

The absolute, city block, or Manhattan metric is

$$d_{ij}=\sum_{k}|X_{ik}-X_{jk}|$$

Its rationale is that if the variables have scale units of equal value, the entities should have the same distance whether two units apart on each of two variables, or one unit apart on one and three on the other.

The "cosine theta" distance is the cosine of the angle between the vector from the origin to point X_{ik} and the vector from the origin to point X_{jk}.

A generalized distance measure is the Minkowski metric:

$$d_{ij}=(\sum_{k}|X_{ik}-X_{jk}|^r)^{1/r}$$

which is a Euclidean metric for r=2 and a city block metric for r=1.

The Mahalonobis distance measure (D^2) is of the form

$$d_{ij}=(X_i-X_j)'\ \Sigma^{-1}(X_i-X_j)$$

where Σ is the pooled-within-groups variance-covariance matrix, and X_i and X_j are the vectors of scores for entities i and j. The Mahalanobis distance allows for correlations between variables; if the variables are uncorrelated, D^2 is equivalent to Euclidean distance measured using standard variables.

The Canberra metric, given below, has the advantage of being unaffected by the range of the variable:

$$d(i,j) = \sum_{k} (|X_{jk} - X_{ik}|)/(X_{ik} + X_{jk})).$$

A modified form, which accommodates negative states, is

$$d(i,j) = \sum_{k} (|X_{jk} - X_{ik}|/(|X_{ik}| + |X_{jk}|)).$$

The Calhoun distance uses only rank orders; for molecules i and j, the distance is the proportion of the entire set (excluding i and j) that have descriptor states intermediate between that for i and that for j for one or more of the descriptors k.

A distance measure may be transformed into a similarity

measure by any of a variety of transformations that convert a non-negative number to the range 0..1, e.g.,

$$S_{ij}=1/(1+d_{ij})$$

A similarity measure may be converted into a distance by, e.g., $d_{ij} = 1-s_{ij}$.

If there is a theoretical maximum distance (d_{tmax}), based on the theoretically possible ranges for each of the component descriptors, the similarity may be expressed as

$$S_{ij}=1-(d_{ij}/d_{tmax})$$

Alternatively, one may calculate the distances between all pairs, and then use the actual maximum distance (d_{amax}):

$$S_{ij}=1-(d_{ij}/d_{amax})$$

Instead of using the ratio of the actual distance to the actual or theoretical maximum distance, one may express s_{ij} as the fraction of the pairs for which the distance is greater than or equal to d_{ij}. This is a measure of relative similarity.

The diversity of a set of compounds, as measured by a set of descriptors, may be calculated in several ways.

A purely geometric method involves assuming that each compound sweeps out a hypersphere in descriptor space, the hypersphere having a radius known as the similarity radius. The total hypervolume in descriptor space of points within a unit similarily radius of one or more of the compounds is calculated. This is compared to the hypervolume achievable if none of hypersphere's overlap; i.e., to n * volume of a single hypersphere, where n is the number of compounds in the set. The swept hypervolume may be determined exactly, or by Monte Carlo methods. The ratio of the swept hypervolume to the maximum hypervolume is a measure of compound set diversity, ranging from 1 (maximum) to 1/n (minimum).

Another approach is to calculate all of the pairwise distances between compounds in descriptor space. The mean distance is a measure of diversity. If desired, this can be scaled by calculating the ratio of the mean distance to the maximum theoretical distance.

A third approach is to apply cluster analysis to the set of compounds. The method used should be one which does not set

the number of clusters arbitrarily, but rather decides the number based on some goodness-of-fit criterion. The resulting number of cluster is a measure of diversity, as is the ratio of the number of clusters to the number of compounds.

One may calculate a measure of disorder for a descriptor as

$$H(k) = - \sum_{g=1}^{m_k} P_{kg} \ln P_{kg}$$

where m_k is the number of different states in descriptor k, and P_{kg} is the observed proportion of individuals exhibiting state g for descriptor k. For uncorrelated descriptors, the sum of H(k) for all k is a measure of overall diversity. Standard techniques may be used to correct for correlation.

Reactivity Descriptors of Protein Binding Sites

Binding sites recognize ligands based on complementarity of both shape and functionality. The functionality in the binding site is an array of recognizable groups (hydrophobic, hydrogen bond donors, hydrogen bond acceptors, π stacking) that is complementary to the ligand (*9*). Many of these functional groups are reactive towards common chemical modification reagents, such as hydroxyl radical, MnO_4^-, $NaBH_4$, and dimethyl sulfate (*10*), and towards "designer" reagents usually based on transition-metal complexes (*11,12*). These reagents can react with functional groups via a wide range of pathways, including one-electron oxidation, oxygen-atom transfer, hydrogen-atom abstraction, hydrogenation, one-electron reduction, hydrolysis, and alkylation. The array of functional groups in a binding site, will therefore exhibit a unique array of rates of reactivity to each of these reagents. This array of rates will provide a set of descriptors for the chemical functionality of the binding site. Therefore, the set of descriptors can be obtained by measuring the relative rates of the functional groups in the binding site against a panel of chemical modification reagents.

It may be the case that most of the solvent accessible functionality is only in the binding site; however, it will be

useful to know that the reactivity descriptors are binding site specific. For example, the structure of lysozyme shows that there are two very important tryptophan residues in the binding site, so it will be important that the measured rate is for those and not for the other trp residues in the protein. On the other hand, the binding site trp residues are by far the most accessible and will probably dominate the measured reactivity. The easiest way to assess this point will be to measure the reaction rates with and without a ligand (such as a peptide BioKey) that blocks the binding site. The difference in rates is then the reactivity of the binding site. Alternatively, the chemical modification reagent can be covalently attached to the ligand and delivered directly to the binding site.

The presently preferred reagents are transition-metal complexes -- e.g., $Pt_2(pop)_4^{4-}$, $Ru(tpy)(bpy)O^{2+}$, and $Ru(bpy)_3^{3+}$ -- whose rate constants can be measured by optical spectroscopy or Stern-Volmer quenching.

Oxoruthenium(IV) complexes. Complexes based on $Ru(tpy)(bpy)O^{2+}$ oxidize DNA by abstraction of the 1′ hydrogen from the deoxyribose ring and by a somewhat more efficient pathway involving inner-sphere oxidation of guanine at C8 (bpy = 2,2′-bipyridine, tpy = 2,2′,2′′-terpyridine) (*18,19*). The oxoruthenium(IV) system has been very useful in discerning the important kinetic and thermodynamic factors in DNA oxidation, because the precise quantity of oxidant is known and the fate of all oxidizing equivalents can be quantitated. Further, the oxidized and reduced forms of the catalyst have distinct optical spectra that allow for the reactions to be followed in real time. These studies have provided insights into the relative reactivities of DNA and RNA toward exogenous oxidants and the relative C-H bond strengths of the sugar hydrogens (*18,19*). In addition, steric factors influencing the efficiencies of inner-sphere guanine oxidation and activation of sugar C-H bonds have been carefully studied (*20*). The reactivity of $Ru(tpy)(bpy)O^{2+}$ toward DNA is summarized in **Figure 1.**

Diplatinum(II) complexes. Generation of hydroxyl radical via the reaction of $Fe(EDTA)^{2-}$ with hydrogen peroxide is a

powerful method for imaging structures of unusual DNA's and DNA-protein complexes (*21*). A parallel approach to the $Fe(EDTA)^{2-}/H_2O_2$ system based on the tetraanionic complex $Pt_2(pop)_4^{4-}$ (pop = $P_2O_5H_2^{2-}$) has been developed (*22-24*). This complex abstracts hydrogen atoms or electrons from organic substrates upon photolysis. In duplex DNA, hydrogens are abstracted from the 4′ and 5′ positions of the sugar, and electrons are abstracted from guanine (*24*). The absolute rate constants for reactions of $Pt_2(pop)_4^{4-}$ can be measured by Stern-Volmer quenching of the emissive excited state (*23*).

Ruthenium(III) complexes. The reactivity pathways available to the oxoruthenium(IV) and $Pt_2(pop)_4^{4-}$ complexes include inner-sphere reactions where there is significant bond-breaking or bond-making in the transition state that involves the metal complex. These are the reaction pathways described above that lead to hydrogen abstraction from nucleotides. Outer-sphere pathways are ones where only an electron is transferred by tunneling from one reactant to another. The advantage of these pathways is that there is a spherically symmetric distance dependence to the reaction probability while inner-sphere pathways generally require a specific approach of the reagent on the reactant. This difference will be important here in defining a diverse set of reaction pathways of the reagents with the protein targets. Complexes based on $Ru(bpy)_3^{3+}$ react with substrates only via outer-sphere electron transfer (*26*). Therefore, unlike $Pt_2(pop)_4^{4-}$, which can oxidized substrates via both inner-sphere hydrogen atom transfer and outer-sphere electron transfer, reactions of $Ru(bpy)_3^{3+}$ complexes will isolate the outer-sphere pathway. The reactions available to $Ru(bpy)_3^{3+}$ will therefore be a subset of those available to $Pt_2(pop)_4^{4-}$. We have used electron-transfer reactions between guanine and $Ru(bpy)_3^{3+}$ as a very sensitive probe of solvent accessibility at guanine in DNA and DNA-protein complexes (*27-29*), which can be used to signal the presence of a single-base mismatch at guanine or base flipping of a paired cytosine into the active site of a DNA repair enzyme. The absolute rate constants for reactions of $Ru(bpy)_3^{3+}$ can be measured by optical spectroscopy, as with $Ru(tpy)(bpy)O^{2+}$.

Reactivity profiles. The amino acid "reactivity profiles"

will be determined for each of the proposed reagents. The rate constants for the reagents may be determined with any or with all 20 conventional amino acids by methods described below. These 20 rate constants will then provide a profile for each reagent. The desired result is that there is a different profile for each reagent. For example, if $Ru(tpy)(bpy)O^{2+}$ reacts mostly with C-H donors like serine and threonine but $Ru(bpy)_3^{3+}$ reacts mostly with one electron donors like tyrosine and tryptophan, then each descriptor and the relationship between the two descriptors will be informative. The rate constants will be measured for the free amino acids and a representative set of di- and tri-peptides to show that the individual rate constants add together linearly and can be discriminated by the three reagents.

The three preferred reagents and their reactivities are well suited to discriminating the 20 amino acids. For example, $Pt_2(pop)_4^{4-}$ abstracts hydrogen atoms from weak C-H bonds in organic substrates. Shown in **Figure 2** are some of the amino acids with the C-H bonds likely to be activated by $Pt_2(pop)_4^{4-}$ highlighted. The C-H activation chemistry of $Ru(tpy)(bpy)O^{2+}$ is distinct from that of $Pt_2(pop)_4^{4-}$ in that inner-sphere adducts of the ruthenium-oxo linkage are often formed (*19,34*), which favors activation of alcohols over aliphatic functionality. Thus, $Ru(tpy)(bpy)O^{2+}$ is likely to prefer serine, threonine, and cysteine more than $Pt_2(pop)_4^{4-}$. Inner-sphere chemistry on the tryptophan and histidine rings or terminal amines of argninine are also likely with $Ru(tpy)(bpy)O^{2+}$. Outer-sphere electron transfer has been shown in proteins for tyrosine, tryptophan, and cysteine (*35,36*), and these reactions should be possibilities with $Ru(bpy)_3^{3+}$. Methionine oxidation at different extents can be expected by all three reagents (*10*). These pathways will be strongly dependent on the folding of the protein (*36*), which should give excellent specificity. The outer-sphere electron transfer reactions will also be observed to some extent with $Pt_2(pop)_4^{4-}$.

The validity of the approach will also be tested using peptides. It is believed that the rate constants for a large array of amino acids will be approximately the average for all of the amino acids separately, weighted by the solvent

accessibility of the individual groups in the folded protein. Therefore, the rate constant for a dipeptide should be the average of the rate constants for each individual amino acid, and similarly to larger peptides as long as no secondary structure develops to attenuate the reactivity of certain groups that are protected by the tertiary structure. It will be instructive to measure the rate constants for some representative random-coil peptides to show how simply linking the amino acids modulates the observed rate constants.

The rate constants may be measured for the three transition-metal complexes under three conditions: denatured protein, folded protein, and folded protein with bound surrogate (BioKey) peptide or nucleic acid. The objective is for the difference in rate constants for folded and denatured protein to give a quantitative descriptor of the surface area of the folded protein (weighted by the solvent-exposed amino acids) and for the rate constants with and without the bound BioKey molecule to give a quantitative descriptor of the amino acids in the binding site. A complication is that the bound BioKey may present new solvent-accessible residues that react with the transition-metal complex. To test for this complication, BioKey peptides with a scrambled amino acid sequence will be included in the reaction with the protein alone, so that the reactive functionality in the BioKey will be present in both reactions.

Unfolded proteins may be generated chemically by addition of a denaturing agent, such as urea or guanidinium hydrochloride; we have shown elsewhere that urea does not deactivate our reagents (*24*). In cases where the chemical denaturant is problematic, we will use thermal denaturation; in the case of $Pt_2(pop)_4^{4-}$, we have shown that thermal denaturation does not alter the selectivity of the reagent with biomolecules except as modulated by the change in biomolecular structure.

The rate constant for the peptide may be determined separately and used to correct for additional oxidation of peptide side chains not blocked by the protein. The rate constant for a peptide with the same amino acids, but with a scrambled sequence, may also be of interest.

The principles of the approach are illustrated in **Figure 3**. In the unfolded protein, all of the amino acids will be

accessible to the transition metal complex. In the folded protein, only the surface residues will be accessible to the reagent. In the ligand-bound folded proteins only the surface residues not occluded by the ligand (i.e., residues outside the binding site) will be accessible to the reagent.

The protein oxidation rates are measured either by quantitating the disappearance of the reagent by standard analytical methods, quantitating the disappearance of the protein by mass spectrometry, amino acid analysis, changes in fluorescence of probes that bind to proteins, or competition with DNA plasmids that are cleaved by the reagent. Rates are compiled for all proteins for each analytical method and each protein state (folded, unfolded, BioKey). Each of these rates is a new quantitative descriptor of the protein.

If one measures the rate of disappearance of the reagent, this will reflect its action on all amino acids. If instead one measures the rate of disappearance of a particular amino acid, or the rate of appearance of the product of the action of the reagent on a particular amino acid, the measurement will be more amino acid-specific. Any or all of these approaches may be used to generate descriptors.

The reactivities may be weighted by the solvent accessibility of each residue as prescribed by the folded structure, so the difference in rates for the folded and unfolded proteins will give a quantitative description of the degree of folding and of the chemical functionality on the solvent-accessible surface. When the BioKey peptide is bound to the active site, the active site residues will be blocked. The difference in the rates with and without the BioKey will therefore be a quantitative descriptor of the number and kind of residues in the active site.

The collected data are used to generate a new database containing *in vitro* descriptors of protein targets. For each protein/reagent combination, the database contains entries for relative rates in the unfolded, folded, and ligand-bound states. The relative rates are normalized on a scale from zero to one and entered into the database in a three-dimensional matrix where each point corresponds to a particular reagent for a given protein in one of the three states. More dimensions may be

added by including amino acid-specific rate measurements, such as surface hydrophobicity changes. In general, proteins that exhibit large changes in rate from unfolded to folded states are those with compact structures and a large number of buried residues. Likewise, large changes in rate between folded and BioKey-bound states indicates large active sites. These trends will be observed as averages across the entire set of reagents. So geometric features will be apparent as averages for *all* reagents. Conversely, differences *between* individual reagents will relate to the composition of the protein surface and active site, because some residues that are protected will be reactive towards some reagents and not others. Thus, the database will provide considerable information on both the geometry AND composition of ligand binding sites (and hence affinity toward particular compounds).

The three reagents described thus far can be generalized by substitution and still be amenable to real-time characterization. For example, the $Ru(tpy)(bpy)O^{2+}$ complexes can be substituted with electron-donating and releasing substituents that will increase or decrease the driving force for oxidation and hence the reactivity and selectivity for different groups in the polymer. For example, we showed that a very electron-rich derivative, $Ru(bpy)_2(4\text{-}NMe_2\text{-}py)O^{2+}$, only abstracts 1′ hydrogens from thymidine sugars and not from A, C, or G whereas the parent $Ru(tpy)(bpy)O^{2+}$ complex oxidizes all four nucleotides (*42*). In addition, these complexes can be substituted with sterically differentiated groups that will modulate the selectivity based on the solvent accessibility. In this vein, we have shown that the steric effects lower the reactivity towards guanine oxidation in duplex DNA without changing the reactivity toward sugar 1′ hydrogens (*20*). Reactivity of one-electron oxidants based on $Ru(bpy)_3^{3+}$ can be modulated by electron-releasing or withdrawing groups or by making the complex significantly larger or smaller, which will change the electron-transfer distance (*27*). Again, these changes should modulate the reactivity profile and the effects of solvent accessibility. Such changes in profile will create reagents that return differentiated descriptors for the protein targets.

The transition-metal reagents are attractive because of the facility with which they can be modified and the ease with which absolute rate constants can be measured (*17*); however, mining the potential information available from the wide range of known chemical modification reagents requires general methods for measuring the relative rates. As discussed earlier, chemical modification reactions have been much more widely applied to study of nucleic acid structure than to study of protein structure because of the greater lability of the phosphodiester backbone. Indeed, reactions that modify DNA nucleotides lead to strand scission if not immediately than almost always after base treatment (*43*). A very sensitive method for measuring DNA modification is by plasmid isomerization (*44,22*). This method is therefore also a sensitive method for measuring the instantaneous concentration of the modification reagent. Therefore, competition of protein targets with plasmid DNA for modification by the exogenous reagent may provide a convenient means for sensitively measuring the relative rate constants. Equally attractive approaches involve detection of modified and unmodified protein by MALDI mass spectrometry (*45*) or detection of unreacted reagent by HPLC with radiolabeling if necessary. The approach described above can be applied to other reagents that modify proteins and nucleic acids, such as dimethyl sulfate, MnO_4^-, $NaBH_4$, and hydroxyl radical. Diversifying the list of reagents will ensure that all twenty of the amino acids are assessed in the resulting descriptors.

A parallel approach described below is to assess the extent of modification of different families of amino acids. For example, the change in surface hydrophobicity of a protein can be assessed by the change in fluorescence of 8-anilino-1-naphthalene-sulfonic acid (ANSA) (*45*). An increase in hydrophobicity measured this way is observed upon protein oxidation, which is very well correlated with formation of tyrosine dimers and oxidation of methionine (*36*). Hydrophobicity and other methods that monitor modification of specific sets or subsets of amino acids, such as changes in reactive carbonyl group or formation of oxidized methionine (*36*), will give descriptors that are distinct from those measured by disappearance of the protein or reagent that give

the reactivity of the entire protein. We can therefore envision more dimensionality in the database where the detection method is diversified and yields a discrete set of amino acid-specific rate constants. Determination of the initial reactivity profiles for each reagent will then be even more important in assessing the difference in these amino acid-specific descriptors for each reagent. By picking the right ensemble of analytical methods, we will obtain different and meaningful descriptors for different modification reagents.

Absolute rate measurements. We have discussed in detail the methods that can be applied to measure the absolute rate constants for reduction of the three families of reagents. These methods are described briefly here.

Optical spectroscopy. The power of the $Ru(tpy)(bpy)O^{2+}$ reagents is that the oxidized and reduced forms of the complexes have very different optical spectra (*46*). Therefore, simple mixing of substrate and catalyst under appropriate conditions and monitoring the optical spectrum of the ruthenium complex allows the absolute rate of ruthenium reduction to be measured. These measurements can be made with a diode array spectrophotometer or, for particularly fast reactions, with a rapid-scanning stopped-flow apparatus (*19*). In the case of complex mechanisms, the returned time-dependent optical spectra can be subjected to factor analysis to detect intermediates and complex reaction pathways (*19,34*). For the purposes here, such detailed mechanistic information need not be pursued. Rather, the method of initial rates will be sufficient for quantitating the reactivity of the amino acids, peptides, and proteins. The one-electron reagents based on $Ru(bpy)_3^{3+}$ can also be analyzed by changes in optical spectra.

Stern-Volmer quenching. In reactions where emissive excited states such as $Pt_2(pop)_4^{4-}$ react with substrates of interest, the absolute rate constants can be measured by Stern-Volmer quenching of the emission (*23*). In this method, the emission of the complex is monitored as a function of concentration of the substrate. A plot of the emission in the absence of the quencher (I°) divided by the emission in the presence of the quencher (I) is linear in the concentration of the quencher ([Q]) according to:

$$I^{\circ}/I = 1 + k_Q \tau [Q] \qquad (1)$$

where τ is the emission lifetime (10 μs) for $Pt_2(pop)_4^{4-}$ and k_Q is the second-order rate constant for the reaction of $Pt_2(pop)_4^{4-}$ with the substrate quencher.

Plasmid Isomerization. The method for measuring the rate constants by competition with plasmid isomerization is shown in **Figure 4**. Reaction of the reagent of interest with supercoiled (Form I) DNA leads to nicking of the DNA to produce a nicked plasmid (Form II), which is readily separated from Form I on an agarose gel (*44,22*). Because the plasmid is as much as 4 kb in length, the plasmid isomerization assay is very sensitive to small amounts of DNA damage. The reaction is then repeated in the presence of the protein for which a relative rate constant is desired. When the protein is modified instead of the plasmid, less isomerization is observed. This assay can then be performed for any reagent that modifies both DNA and amino acids.

A drawback to the approach in **Figure 4** is the that the assay cannot be performed for DNA-binding proteins; however, the simplicity and low quantities of material required are attractive. A more general strategy would be to analyze the oxidized protein by matrix-assisted laser-desorption ionization (MALDI) mass spectrometry (*45*). This method can be used to detect small changes in molecular weight of large proteins on 0.1 pmol of protein (for a recent example, see (*45*)). In the simplest case, the mass spectrometry will simply be used to detect the change in the concentration of the unmodified protein as a means for determining the rate of modification by the reagent. A final general approach to measuring the relative rates would be to measure the quantity of reagent before and during reaction with the protein. This method could involve traditional analytical chemistry techniques such as HPLC or GC for suitable substrates. The use of isotopically labeled reagents will provide the desired sensitivity.

As discussed above, a separate strategy for measuring the rates would be rather than to measure the total rate for all of the amino acids, to use detection methods that sample a subset of the amino acids. Such assays would involve quantitation of

reactive carbonyl in the oxidized protein following the reaction, analysis for individual oxidized amino acids such as methionine sulfoxide, or measuring the change in hydrophobicity by ANSA emission (*36*). These methods would then provide further specificity in the descriptors that would be meaningful when the same detection methods were compared for different proteins.

Individual protein information. For each crystallographically characterized protein and reagent, the reactivity desciptors should be predictable from the three-dimensional structure and the reactivity profile, which describes the inherent chemical reactivity of each amino acid towards the reagent. Thus, the three-dimensional structure can be used to weight each amino acid by its solvent accessibility in the folded protein, which can then be weighted by its inherent reactivity. The ability to perform these calculations carefully will depend on accurate reactivity profiles, which is why choosing transition-metal reagents where absolute rate constants can be measured is vital in the initial studies. When rates are measured in the presence of the BioKey peptide, again the reactivity should be predictable from the three-dimensional structure and the reactivity profile.

The amino acid composition of protein binding sites and surfaces has been determined in 50 proteins whose crystal structures with bound ligands are known (*9*). These studies show that Trp, His, Arg, and Tyr are much more abundant in binding sites contacting bound ligands than in general in the protein. Gly and Ser are often found near the bound ligand as well; however, these residues are generally abundant throughout the protein. Therefore, reagents that are specific for Trp and His, which are found at very low frequencies outside the binding site, will be very informative. Attractive reagents include $Ru(tpy)(bpy)O^{2+}$, which oxidizes ring nitrogens, reagents that alkylate ring nitrogens, or oxidants that form amine oxides at ring nitrogens. *Importantly, the distribution of amino acids found on protein surfaces is fairly similar to the distribution found on average, whereas the distribution found in binding sites is dramatically different from the average* (*9*). Thus, residues protected by binding of the BioKey ligand will have a very different profile from those protected by folding of the

protein.

Database construction.

The collected data will be used to generate a new database containing *in vitro* descriptors of protein targets, which we will call **R** (for reactivity or recognition) to distinguish it from NCI's **S**, **A**, and **T** databases.

In a preferred embodiment, for each protein and each reagent (or reaction condition, if a particular reagent is used under several different reaction conditions), the database preferably contains information regarding reaction rates in the unfolded, folded, and ligand-bound states. Rate constants may provided for several different ligand-bound states, i.e., with different ligands. Preferably, the rate information is expressed in relative terms, and normalized on a scale from zero to one. The rate information may be expressed in difference form, e.g., by providing the (rate folded-rate unfolded), (rate ligand bound-rate folded), (rate ligand bound-rate unfolded), and/or (rate ligand 1 bound-rate ligand 2 bound).

The data may be conceptualized as a three-dimensional matrix where each point corresponds to a particular reagent for a given protein in one of the three states, although its most efficient representation is likely to be that of a relational database.

In a preferred embodiment, the various databases are normalized in accordance with standard database programming practice to minimize the duplication of information.

Thus, one database may be of reactions. Each record in this database will contain a reaction ID, and will normally contain additional information about the reaction, such as the chemical name of the reagent, the reaction conditions (e.g., temperature, solvent), and assay method. If a reagent is used under many different reaction conditions, is may be appropriate to also create a reagent database, with a reagent ID field used to link the reaction and reagent databases, and information about the reagent placed in the other fields of the reagent database record.

Another database may tabulate target proteins. Each record in this database will contain a target ID, and may optionally

contain additional information about the target protein, such as its name, biological activity, sequence, etc. A third database could provide the target protein state. Each record of the "state" database will have a state ID field, and additional fields which identify whether the state is a folded or unfolded protein and, if folded, whether a ligand is bound to it. A ligand could be identified by a ligand ID acting as a lookup field for retrieval of information from a ligand database.

The reaction, protein, and protein state databases are relationally linked to the reactivity database, with the reaction ID used for lookup of a reaction and the target protein ID used for lookup of a target protein. Thus, each record in the reactivity database will contain fields for the reaction ID, the target protein ID, and the result of the reaction (e.g., a rate constant).

There is a many-to-one relationship between the reaction database and the reactivity database, and a one-to-many relationship between the reactivity database and the protein database.

Likewise, there is a database of drugs which interact with one of more of the proteins. Since there is a many-to-many relationship between the drugs and the proteins, this relationship is preferably normalized by means of another database of individual drug-protein interactions.

It is also desirable to have an assay database, in which each record identifies a different type of assay. Proteins may have more than one activity, and have a different spectrum of relevant drugs for each.

Therefore, in a preferred interaction database, each record will include a drug ID, a protein ID, an assay ID, and an assay value or potency. Each ID field is a relational link to another database.

It will be appreciate by those skilled in the database programming art that there are many ways of encoding the information of interest. Therefore, the present invention is not limited to any particular database design.

Use of the Reactivity Database

Once the R database is compiled, it will be used to improve the efficiency of lead generation as hereafter described. As discussed above, the initial database will be compiled for proteins that bind known inhibitors and have well understood biology and expression profiles. Once a sufficient number of these proteins are in the database, reactivity descriptors may be determined for proteins that have not been screened against combinatorial libraries, whose three-dimensional structures are not known, and/or whose biological profile has not yet been determined. It will be appreciated that there is no fixed minimum number of proteins which must be in the database before it can provide useful information about a protein of interest. The more database proteins there are with similar reactivity descriptors, and the more similar those descriptors are, the more useful the prior knowledge of ligands for those database proteins is likely to be. Plainly, the odds improve as the number of proteins in the database increases. Preferably, the database will provide data on at least 50, more preferably at least 200, still more preferably at least 1000, and most preferably every known protein. Another factor is the diversity of the proteins in the database. The more diverse the proteins, the more likely it is that at least one database protein will have a reasonably similar reactivity.

The reactivity descriptor for a protein is the set of rate constants (or rate constant differences) for all of the characterizing reactions to which that protein has been subjected. The reactivity descriptors will be entered into R and the similarity between the unknown protein and the known proteins in the database will be determined as described in the section on "Descriptors".

The utility of a reactivity database is dependent, not only on the number of proteins in the database, but also on the number of reactions to which the protein was subjected, and the number of protein states examined. All else being equal, the more data points, the greater the degree of characterization. Preferably, there are at least 1, more preferably at least 3, still more preferably at least 10, most preferably at least 100, datapoints (target protein/state/reaction triplets). While less easily defined, the greater the diversity in the

characterizing reactions, the more useful the database will be.

Relation to databases with information on activity (such as **A**) will predict what types of chemical compounds will most likely bind to the unknown target, because proteins that appear similar in **R** will bind similar compounds. This procedure will be analogous to calculating similarities between chemical compounds (such as in **S**) and predicting that compounds similar to a known inhibitor will also inhibit similar targets. Relation to an expression database such as **T** will provide pharmacological information on the unknown target.

Aptamer-Based Descriptors of Protein Binding Sites

An aptamer-based descriptor is a description of a protein in terms of the aptamers which recognize it. While peptides could serve as aptamers, the preferred aptamers are nucleic acids.

Such a descriptor is not a simple numerical value. Rather, it is a list of sequences (and, preferably, secondary structures and contact points) for each of the aptamers identified as binding a particular protein. In this section, we will describe how nucleic acid aptamers are identified and characterized, and how the similarity of the aptamer-based descriptors for two different proteins may be calculated.

In essence, a single-stranded oligonucleotide library is screened to identify aptamers which bind a protein with a desired affinity. This protein may be a reference protein with known drug antagonists, or the target protein for which such antagonists are to be identified. The desired aptamers are amplified and sequenced.

The aptamers serve to characterize the protein in that only those oligonucleotides which can conform to the surface of the protein will bind to it. One may say that the aptamers take "impressions" of the protein surface. Some of the aptamers may be expected to bind the protein at a site corresponding or, overlapping, or otherwise occluding the functional site of the protein. (Such aptamers may, if desired, be identified by screening the aptamers for antagonist activity.) Others will bind at sites distal to the functional site. Some will bind to the same site, others, to different sites. All contribute to

a "picture" of the protein.

Preferably, the contact sites (the bases within the aptamer) through which the aptamer contacts the protein are identified. A preferred means for such identification is a footprinting reaction where chemical modification of the nucleic acid is performed with and without the bound protein; sites where bound protein blocks chemical modification are contact sites. Footprinting of nucleic acids in this manner can be achieved using enzymes such as DNAse or chemical reagents such as copper-phenanthroline (Papavassiliou, A.G. Biochem. J. 1995, 305, 345-357), $Fe(EDTA)^2$ (Pogozelski, et al., J. Am. Chem. Soc., 117:6428-33, 1995), or $Pt_2(pop)_4^4$ (Breiner, K.M.; Daugherty, M.A.; Oas, T.G.; Thorp, H.H. J. Am. Chem. Soc. 1995, 117, 11673-11679).

Preferably, after the contact sites have been determined, the secondary structure of at least the contacting bases of the oligonucleotide is analyzed. The secondary structures of the aptamers can be predicted using the approach of Zichi et al. (J.P. Davis, N. Janjic, D. Pribnow, D.A. Zichi, Nucleic Acids Res. 1995, 23, 4471-4479) where a two-dimensional color grid is used to indicate sites of potential base pairing, revealing the underlying secondary structure. More traditional nucleic acid folding approaches are those of Zuker, which can be found in M. Zuker, Science 1989, 244, 48-52 and J. Jaeger, D. Turner, and M. Zuker, Proc. Natl. Acad. Sci USA 1989, 86, 7706-7710. The contact sites, determined using the footprinting reactions described above, are mapped onto the predicted secondary structure, and the functionality sequence is the list of nucleotides that contact the protein, read from the 5' to 3' direction. The secondary structure may also be determined experimentally, see Tinoco, Jr., J. Phys. Chem., 100:13311-22 (1996).

Thus, the descriptor for each aptamer preferably identifies not only the overall sequence of the aptamer, but also the contact site. In the cases of an oligonucleotide, the bases of the contact site are preferably described not only by identifying the base itself, but also indicating whether in the secondary structure of the aptamer it is paired to any base, and if so what. In a convenient notation to indicate the secondary

structure at each contact site, two-letter codes are used for each nucleotide. For single-stranded nucleotides, the contact nucleotides are followed by the small letter "o", so the contacts are Ao, To, Go, and Co for a DNA aptamer. For double-stranded nucleotides, the contact sites are followed by a small letter representing the nucleotide on the opposite strand. For example, if the protein contacts an A in an AT base pair from cleavage, that site is listed as "At". Both base pairs and mismatches are represented this way, so the entire list of double-stranded codes is (for DNA): At, Aa, Ag, Ac, Gc, Gg, Gt, Ga, Cg, Cc, Ct, Ca, Ta, Tg, Tt, Tc. When combined with the single-stranded codes, there are 20 possible functionality elements.

An alternative to comparing the footprinted sequence in a linear array would be to develop a two-dimensional projection of the secondary structure of the aptamer. For example, suppose the aptamer at the right is selected for a given target and the sites indicated with an arrow are determined to be protected by the target via Pt-pop footprinting. The aptamer can then me mapped onto a two-dimensional grid where the contact sites are coded as before and the remaining "placeholder" sites are coded as either Nn for a base pair or mismatch site or No for a single-stranded site (see **Figure 5**). There is no need to differentiate the base pair and mismatch placeholder sites because this information is already in the contact site codes. The two-dimensional grid can now be analyzed for similarity to other two-dimensional representations by graph-theoretical approaches, such as those used for determining compound similarity (as in Patterson et al., J. Med. Chem., 1996).

The contact sites and secondary structure are used to develop a consensus "functionality sequence" (epitope), which represents both the primary sequence and the secondary structure of the nucleic acid moiety which contacts the target protein. These functionality sequences are entered into a relational database that is used to analyze functionality sequences for unknown targets. Targets with homologous functionality sequences bind small molecules of a similar nature. Thus, if a connection is drawn between an unknown target and a target for which small molecule binders are known, screening of the unknown

target can be restricted initially to libraries of molecules similar to those that bind the known target. Functionality sequences can be determined on small quantities of targets, and the knowledge of the physiological function of the unknown target is not required. Related strategies involving combinatorial peptides and determination of contact sites using mutagenesis can also be envisioned.

Sequence identity among aptamer contact sites may be determined using an adaptation of BLAST from the National Center for Biotechnology Information. The BLAST (Basic Local Alignment Search Tool) algorithm is described in S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, J. Mol. Biol. 215, 403-10 (1990). A new version that allows the functionality (secondary structure) codes (At, Ta, Ao, etc.) to be compared can be adapted from art in the public domain.

In the simplest comparison scheme, an identity matrix is used. If the aligned functionality elements are the same, the score for that pair of elements is one, if they are different, the scope is zero. The individual scores are summed and divided by the total number of elements.

A second approach would be to weight more highly the alignments least likely to occur by chance alone. This would require tabulating predicted secondary structure information for a large number of DNA sequences. This could be done by (1) randomly generating DNA sequences of the length and base composition to be used in the library, (2) predicting the secondary structure of each sequence by standard methods, (3) converting the string of bases into a string of "functionality elements" and (4) calculating the probability of occurrence of each alignment. Of course, this weighting scheme is not limited to protein contact sites; all bases are considered.

For example, a base pair (Gc, Cg, At, Ta) is likely to be more common, and hence less informative, than a mismatch (Gt, Gg, Ga, Ag, etc.) or a single-stranded nucleotide (Go, Ao, To, Co). These elements could therefore be weighted appropriately to allow the elements that have the greatest information to be considered most heavily in determining similarity. For one possible weighting, see Ex. 3.

While nucleic acids are preferred, because they can be

amplified, peptides may also be used to generate aptamer-based descriptors. Preferably, these peptides are 5-10 a.a. in length. The peptide library is screened for binding to the query protein and the peptide aptamers are compared to those of the peptide aptamers found to bind each of the reference proteins. BLAST may be used in its standard form to compare the amino acid sequences. Any scoring matrix conventional in the art, e.g., BLOSUM, may be used to score aligned amino acid pairs. The resulting sequence similarity score may then be standardized.

Instead of determining aptamer similarity on the basis of sequence similarity, it may be determined using an adaptation of structural descriptors. For example, in the case of nucleic acid aptamers, a 2-D fingerprint may be devised in which each bit represents the presence or absence of a particular secondary structure, such as an unbonded region, an unbonded region of a particular length (e.g., 2-6 bases, 7-20 bases, >20 bases), an interior loop, an interior loop of a particular length (e.g., 2-6 bases, 7-20 bases, >20 bases), a bulge loop, a bulge loop of a particular length (e.g., 1 base, 2-3 bases, 4-7 bases, 8-20 bases, >20 bases), a hairpin loop of any length and type, a hairpin loop closed by G:C, a hairpin loop closed by A.U, a particular type of hairpin loop of a particular length (e.g., 3 bases, 4-5 bases, 6-7 bases, 8-9 bases, 10-30 bases, >30 bases), a run of paired bases of a particular length, an overall base composition in a particular range (e.g., 40-60% GC), etc. See Tinoco, et al., Nature New Biol., 246:40-41 (1973) and Tinoco, et al., Nature, 230:362-7 (1971) for an early approach to estimating RNA structure which is readily adapted to 2D fingerprinting. One can readily build on this work, e.g., add a bit representing the presence or absence of the extrastable tetraloop 5'-GGAC(UUCG)GUCC 3', or one for a higher structure such as a pseudoknot. Particular consideration may be given to structures often implicated in single stranded nucleic acid:protein binding.

Suppose that m aptamers bind the query protein, and n aptamers binding a reference protein. The aptamer-based similarity of the target and reference protein may be calculated on the basis of any or all of the m x n possible comparisons.

Many approaches are possible, including, but not limited to

1. comparing the highest affinity query protein-binding aptamer with the highest affinity references protein binding aptamer.

2. Making all (m x n) comparisons and average the results.

3. Calculating a weighted average of the m x n comparisons, weighted on the basis of the affinities. The highest affinity query-binding aptamer is given a relative affinity of 1, and the other aptamers are given suitable relative affinities (which will be fractions of 1). These relative affinities may be assigned on the basis of the actual affinities, or their loganthms. The same is done with the reference-binding aptamers. The similarity score for each of the m x n comparisons is then weighted by the product of the mth query-binding relative affinity and the nth reference-binding relative affinities.

4. Developing a consensus sequence for the m query-binding aptamers and a consensus sequence for the n reference binding aptamers. The two consensus sequences are then examined.

5. Making all (m x n) comparisons and select the highest score.

6. Making all (m x n) comparisons and select the lowest score.

7. Making all (m x n) comparisons and select the median score.

After functionality sequences have been determined for a sufficient number of known targets, aptamers are selected for new targets of unknown structure. The ligands for these targets are footprinted, e.g., using $Pt_2(pop)_4^{4-}$, and the functionality sequences are determined. The functionality sequences are entered into the data base, and connections are drawn between the ligands that bind the unknown target and those already in the data base for known targets. Initial screening of small molecule libraries is then directed towards compounds that have similar size and functionality to those that bind the known targets that exhibit high functionality sequence homology to the unknown target of interest.

In one embodiment, this information is entered into a set of relationally linked databases. We have already discussed

protein, drug, and protein-drug interaction databases. Clearly, we may also provide a database listing all aptamers which bind at least one protein in the database. The aptamers database has a many-to-many relationship with the protein database. Hence, to normalize this relationship, we would like an aptamer-protein interaction database, in which each record includes an aptamer ID field (a link to an aptamer database) and a protein ID field (a link to a protein database). The record may optionally identify the contact site. Secondary structure may be indicated in either the aptamer record (if invariant) or in the aptamer-protein interaction record (if affected by the protein binding).

Other Protein Descriptors

In addition to the chemical reactivity and aptamer binding descriptors discusses above, the database may include additional descriptors for the target proteins. Possible descriptors include the following:

- Amino Acid Composition
- Structure
  - Predicted Overall Alpha Helicity
  - Predicted Number of Alpha Helices or Beta Strands
  - Disulfide Bond Topology
  - Predicted Surface Area-to-Volume Ratio
- Sequence
  - Similarity Scores vis-a-vis selected Reference Proteins
- Gross Physical Characteristics
  - Molecular Weight
  - Isoelectric Point
  - Thermostability
  - Overall Hydrophobicity
  - Overall Aromaticity
  - CHO content
- Biological Activity (various)

Compound Library

The compound library is a combinatorial library whose members are suitable for use as drugs if, indeed, they have the ability to mediate a biological activity of the target protein.

Peptides have certain disadvantages as drugs. These include susceptibility to degradation by serum proteases, and difficulty in penetrating cell membranes. Preferably, all or most of the compounds of the compound library avoid, or at least do not suffer to the same degree, one or more of the pharmaceutical disadvantages of peptides.

In designing a compound library, it is helpful to bear in mind the methods of molecular modification typically used to obtain new drugs. Three basic kinds of modification may be identified: disjunction, in which a lead drug is simplified to identify its component pharmacophoric moieties; conjunction, in which two or more known pharmacophoric moieties, which may be the same or different, are associated, covalently or noncovalently, to form a new drug; and alteration, in which one moiety is replaced by another which may be similar or different, but which is not in effect a disjunction or conjunction. The use of the terms "disjunction", "conjunction" and "alteration" is intended only to connote the structural relationship of the end product to the original leads, and not how the new drugs are actually synthesized, although it is possible that the two are the same.

The process of disjunction is illustrated by the evolution of neostigmine (1931) and edrophonium (1952) from physostigmine (1925). Subsequent conjunction is illustrated by demecarium (1956) and ambenonium (1956).

Alterations may modify the size, polarity, or electron distribution of an original moiety. Alterations include ring closing or opening, formation of lower or higher homologues, introduction or saturation of double bands, introduction of optically active centers, introduction, removal or replacement of bulky groups, isosteric or bioisosteric substitution, changes in the position or orientation of a group, introduction of alkylating groups, and introduction, removal or replacement of groups with a view toward inhibiting or promoting inductive (electrostatic or conjugative (resonance) effects.

Thus, the substituents may include electron acceptors and/or electron donors. Typical electron donors (+I) include $-CH_3$, $-CH_2R$, $-CHR_2$, $-CR_3$ and $-COO^-$. Typical electron acceptors (-I) include $-NH_3+$, $-NR_3+$, $-NO_2$, -CN, -COOH, -COOR, -CHO, -COR,

-COR, -F, -Cl, -Br, -OH, -OR, -SH, -SR, $-CH=CH_2$, $-CR=CR_2$, and -C=CH.

The substituents may also include those which increase or decrease electronic density in conjugated systems. The former (+R) groups include $-CH_3$, $-CR_3$, -F, -Cl, -Br, -I, -OH, -OR, -OCOR, -SH, -SR, $-NH_2$, $-NR_2$, and -NHCOR. The later (-R) groups include $-NO_2$, -CN, -CHC, -COR, -COOH, -COOR, $-CONH_2$, $-SO_2R$ and $-CF_3$.

Synthetically speaking, the modifications may be achieved by a variety of unit processes, including nucleophilic and electrophilic substitution, reduction and oxidation, addition elimination, double bond cleavage, and cyclization.

For the purpose of constructing a library, a compound, or a family of compounds, having one or more pharmacological activities (which need not be related to the known or suspected activities of the target protein), may be disjoined into two or more known or potential pharmacophoric moieties. Analogues of each of these moieties may be identified, and mixtures of these analogues reacted so as to reassemble compounds which have some similarity to the original lead compound. It is not necessary that all members of the library possess moieties analogous to all of the moieties of the lead compound.

The design of a library may be illustrated by the example of the benzodiazepines. Several benzodiazepine drugs, including chlordiazepoxide, diazepam and oxazepam, have been used on anti-anxiety drugs. Derivatives of benzodiazepines have widespread biological activities; derivatives have been reported to act not only as anxiolytics, but also as anticonvulsants, cholecystokinin (CCK) receptor subtype A or B, kappa opioid receptor, platelet activating factor, and HIV transactivator Tat antagonists, and GPIIbIIa, reverse transcriptase and ras farnesyltransferase inhibitors.

The benzodiazepine structure has been disjoined into a 2-aminobenzophenone, an amino acid, and an alkylating agent. See Bunin, et al., Proc. Nat. Acad. Sci. USA, 91:4708 (1994). Since only a few 2-aminobenzophenone derivatives are commercially available, it was later disjoined into 2-aminoarylstannane, an acid chloride, an amino acid, and an alkylating agent. Bunin, et al., Meth. Enzymol., 267:448 (1996). The arylstannane may

be considered the core structure upon which the other moieties are substituted, or all four may be considered equals which are conjoined to make each library member.

A basic library synthesis plan and member structure is shown in Figure 1 of Fowlkes, et al., U,S. Serial No. 08/740,671, incorporated by reference in its entirety. The acid chloride building block introduces variability at the R^1 site. The R^2 site is introduced by the amino acid, and the R^3 site by the alkylating agent. The R^4 site is inherent in the arylstannane. Bunin, et al. generated a 1, 4-benzodiazepine library of 11,200 different derivatives prepared from 20 acid chlorides, 35 amino acids, and 16 alkylating agents. (No diversity was introduced at R^4; this group was used to couple the molecule to a solid phase.) According to the Available Chemicals Directory (HDL Information Systems, San Leandro CA), over 300 acid chlorides, 80 Fmoc-protected amino acids and 800 alkylating agents were available for purchase (and more, of course, could be synthesized). The particular moieties used were chosen to maximize structural dispersion, while limiting the numbers to those conveniently synthesized in the wells of a microtiter plate. In choosing between structurally similar compounds, preference was given to the least substituted compound.

The variable elements included both aliphatic and aromatic groups. Among the aliphatic groups, both acyclic and cyclic (mono- or poly-) structures, substituted or not, were tested. (While all of the acyclic groups were linear, it would have been feasible to introduce a branched aliphatic). The aromatic groups featured either single and multiple rings, fused or not, substituted or not, and with heteroatoms or not. The secondary substitutents included $-NH_2$, -OH, -OMe, -CN, -Cl, -F, and -COOH. While not used, spacer moieties, such as -O-, -S-, -OO-, -CS-, -NH-, and -NR-, could have been incorporated.

Bunin et al. suggest that instead of using a 1, 4-benzodiazepine as a core structure, one may instead use a 1, 4-benzodiazepine-2, 5-dione structure.

As noted by Bunin et al., it is advantageous, although not necessary, to use a linkage strategy which leaves no trace of the linking functionality, as this permits construction of a

more diverse library.

Other combinatorial nonoligomeric compound libraries known or suggested in the art have been based on carbamates, mercaptoacylated pyrrolidines, phenolic agents, aminimides, N-acylamino ethers (made from amino alcohols, aromatic hydroxy acids, and carboxylic acids), N-alkylamino ethers (made from aromatic hydroxy acids, amino alcohols and aldehydes) 1, 4-piperazines, and 1, 4-piperazine-6-ones.

DeWitt, et al., Proc. Nat. Acad. Sci. (USA), 90:6909-13 (1993) describes the simultaneous but separate, synthesis of 40 discrete hydantoins and 40 discrete benzodiazepines. They carry out their synthesis on a solid support (inside a gas dispersion tube), in an array format, as opposed to other conventional simultaneous synthesis techniques (e.g., in a well, or on a pin). The hydantoins were synthesized by first simultaneously deprotecting and then treating each of five amino acid resins with each of eight isocyanates. The benzodiazepines were synthesized by treating each of five deprotected amino acid resins with each of eight 2-amino benzophenone imines.

Chen, et al., J. Am. Chem. Soc., 116:2661-62 (1994) described the preparation of a pilot (9 member) combinatorial library of formate esters. A polymer bead-bound aldehyde preparation was "split" into three aliquots, each reacted with one of three different ylide reagents. The reaction products were combined, and then divided into three new aliquots, each of which was reacted with a different Michael donor. Compound identity was found to be determinable on a single bead basis by gas chromatography/mass spectroscopy analysis.

Holmes, USP 5,549,974 (1996) sets forth methodologies for the combinatorial synthesis of libraries of thiazolidinones and metathiazanones. These libraries are made by combination of amines, carbonyl compounds, and thiols under cyclization conditions.

Ellman, USP 5,545,568 (1996) describes combinatorial synthesis of benzodiazepines, prostaglandins, beta-turn mimetics, and glycerol-based compounds. See also Ellman, USP 5,288,514.

Summerton, USP 5,506,337 (1996) discloses methods of preparing a combinatorial library formed predominantly of

morpholino subunit structures.

Heterocylic combinatorial libraries are reviewed generally in Nefzi, et al., Chem. Rev., 97:449-472 (1997). One or more moieties of the following types may be incorporated into compounds of the library, as many drugs fall into one or more of the following categories:

acetals

acids

alcohols

amides

amidines

amines

amino acids

amino alcohols

amino ethers

amino ketenes

ammonium compounds

azo compounds

enols

esters

ethers

glycosides

guanidines

halogenated compounds

hydrocarbons

ketones

lactams

lactones

mustards

nitro compounds

nitroso compounds

organo minerals

phenones

quinones

semicarbazones

stilbenes

sulfonamides

sulfones

thiols

thioamides

thioureas

ureas

ureides

urethans

Without attempting to exhaustively recite all pharmacological classes of drugs, or all drug structures, one or more compounds of the chemical structures listed below have been found to exhibit the indicated pharmacological activity, and these structures, or derivatives, may be used as design elements in screening for further compounds of the same or different activity. (In some cases, one or more lead drugs of the class are indicated.)

hypnotics

- higher alcohols (clomethiazole)
- aldehydes (chloral hydrate)
- carbamates (meprobamate)
- acyclic ureides (acetylcarbromal)
- barbiturates (barbital)
- benzodiazepine (diazepam)

anticonvulsants

- barbiturates (phenobarbital)
- hydantoins (phenytoin)
- oxazolidinediones (trimethadione)

succinimides (phensuximide)
acylureides (phenacemides)

narcotic analgesics
morphines
phenylpiperidines (meperidine)
diphenylpropylamines (methadone)
phenothiazihes (methotrimeprazine)

analgesics, antipyretics, antirheumatics
salicylates (acetylsalicylic acid)
p-aminophenol (acetaminophen)
5-pyrazolone (dipyrone)
3, 5-pyrazolidinedione (phenylbutazone)
arylacetic acid (indomethacin)
adrenocortical steroids (cortisone, dexamethasone,
prednisone, triamcilone)
athranilic acids

neuroleptics
phenothiazine (chlorpromazine)
thioxanthene (chlorprothixene)
reserpine
butyrophenone (halopendol)

anxiolytics
propandiol carbamates (meprobamate)
benzodiazepines (chlordiazepoxide, diazepam, oxazepam)

antidipressants
tricyclics (imipramine)

muscle/relaxants
propanediols and carbamates (mephenesin)

CNS stimulants
xanthines (caffeine, theophylline)
phenylalkylamines (amphetamine)
(Fenetylline is a conjunction of theophylline and

amphetamine)
oxazolidinones (pemoline)

cholinergics
choline esters (acetylcholine)
N,N-dimethylcarbamates

adrenergics
aromatic amines (epinephrine, isoproterenol, phenylephrine)
alicyclic amines (cyclopentamine)
aliphatic amines (methylhexaneamine)
imidazolines (naphazoline)

anti-adrenergics
indolethylamine alkaloids (dihydroergotamine)
imidazoles (tolazoline)
benzodioxans (piperoxan)
beta-haloalkylamines (phenoxybenzamine)
dibenzazepines (azapetine)
hydrazinophthalazines (hydralazine)

antihistamines
ethanolamines (diphenhydramine)
ethylenediamines (tripelennomine)
alkylamines (chlorpheniramine)
piperazines (cyclizine)
phenothiazines (promethazine)

local anesthetics
benzoic acid
esters (procaine, isobucaine, cyclomethycaine)
basic amides (dibucaine)
anilides, toluidides, 2, 6-xylidides (lidocaine)
tertiary amides (oxetacaine)

vasodilators
polyol nitrates (nitroglycerin)

diuretics

xanthines
thiazides (chlorothiazide)
sulfonamides (chlorthalidone)

antihelmintics
cyanine dyes

antimalarials
4-aminoquinolines
8-aminoquinolines
pyrimidines
biguanides
acridines
dihydrotriazines
sulfonamides
sulfones

antibacterials
antibiotics
penicillins
cephalosporins
octahydronapthacenes (tetracycline)
sulfonamides
nitrofurans
cyclic amines
naphthyridines
xylenols

antitumor
alkylating agents
nitrogen mustards
aziridines
methanesulfonate esters
epoxides
amino acid antagonists
folic acid antagonists
pyrimidine antagonists
purine antagonists

antiviral
adamantanes
nucleosides
thiosemicarbazones
inosines
amidines and guanidines
isoquinolines
benzimidazoles
piperazines

For pharmacological classes, see, e.g., Goth, Medical Pharmacology: Principles and Concepts (C.V. Mosby Co.: 8th ed. 1976); Korolkovas and Burckhalter, Essentials of Medicinal Chemistry (John Wiley & Sons, Inc.: 1976). For synthetic methods, see, e.g., Warren, Organic Synthesis: The Disconnection Approach (John Wiley & Sons, Ltd.: 1982); Fuson, Reactions of Organic Compounds (John Wiley & Sons: 1966); Payne and Payne, How to do an Organic Synthesis (Allyn and Bacon, Inc.: 1969); Greene, Protective Groups in Organic Synthesis (Wiley-Interscience). For selection of substituents, see e.g., Hansch and Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology (John Wiley & Sons: 1979).

The library is preferably synthesized so that the individual members remain identifiable so that, if a member is shown to be active, it is not necessary to analyze it. Several methods of identification have been proposed, including:

(1) encoding, i.e., the attachment to each member of an identifier moiety which is more readily identified than the member proper. This has the disadvantage that the tag may itself influence the activity of the conjugate.

(2) spatial addressing, e.g., each member is synthesized only at a particular coordinate on or in a matrix, or in a particular chamber. This might be, for example, the location of a particular pin, or a particular well on a microtiter plate, or inside a "tea bag".

The present invention is not limited to any particular form of identification.

However, it is possible to simply characterize those members of the library which are found to be active, based on

the characteristic spectroscopic indicia of the various building blocks.

Solid phase synthesis permits greater control over which derivatives are formed. However, the solid phase could interfere with activity. To overcome this problem, some or all of the molecules of each member could be liberated, after synthesis but before screening.

Lead Generation

As a result of querying the database with descriptor data for a query protein, the user receives a list of reference proteins. For each reference protein a similarity score, and a list of known antagonists (or other modulators) is given. These are the "drug leads". The antagonists are weighted by the similarity scores of their respective proteins. If available, and desired, they may also be weighted by their potency against their corresponding reference protein, and/or by other physicochemical characteristics of interest, e.g., lipophilicity.

This invention contemplates the construction of a composite combinatorial compound library which is biased in favor of compounds (both scaffoldings and substituents) which are structurally similar to the drug leads.

Preferably, in view of the wide variety of "drug-type" compound combinatorial libraries already available, the composite library is based on these already available simple libraries.

Each drug lead is compared, using structural descriptors, to the basic scaffold (or to the most similar member) of each candidate simple combinatorial library. The structural descriptors which may be used include, but are limited to, those listed in Patterson, et al. (1996), Klebe and Abraham (1993), Cummins, et al. (1996), and Matter (1997). Conventional mathematical methods may be used to select or weight the descriptors.

The 2D fingerprint method described in Matter, et al. (1997) is of particular interest. In essence, the compound is analyzed for the presence or absence of particular molecular fragments, the results being encoded in a binary format. In

order to encode the status of a large number of fragments, without assigning a bit to each fragment, each fragment was projected using a pseudorandomization algorithmin to a bitstring of limited size (i.e., fewer bits than the total number of unique fragments in the compounds of the database). In addition, the presence of 60 specific functional groups, rings or atoms was encoded in 60 of the total 988 bits. Details are given in UNITY Chemical Information Software, version 2.5. Reference Guide, pp. 45-58, Tripos Inc., 1699 Hanley Rd., St. Louis, MO 63144.

A similar approach is described in Martin, et al., J. Med. Chem., 38:1431-6 (1995). A "Daylight fingerprint" routine was used to search a molecule for all substructures up to seven bonds long and set one bit in a 2048-bit string for each fragment found. A "hashing" algonthm randomly assigned each fragment to one of the possible bits.

The presence or absence of the molecular frameworks identified by Bemis and Murcko, J. Med. Chem., 39:2887-93 (1996) may be usefully incorporated into binary descriptors of a 2D fingerprint-type.

Of course, other structural descriptors may be used instead of or in addition to 2D fingerprints.

The candidate library is assigned a weight which is a function of (1) the drug lead's query score (reflecting the similarity of its reference protein to the query protein and, optionally, the potency or other drug characteristics of the drug lead) and (2) the structural similarity score between the drug lead and the scaffold. These weights then determine the predominance of that candidate library in the ideal composite library. Thus, if benzodiazepines score twice as high as carbamates, the benzodiazepine library screened might be twice as big as the carbamate library.

Each drug lead may also be used to evaluate possible substituents to be conjugated to the scaffold. Each candidate substituent is scored, on the basis of structural descriptors, for its similarity to the drug lead. The proportion of the candidate substituent in the combinatorial reaction mix may then be governed by its similarity score. Low-scoring candidate substituents may be omitted entirely, or merely reduced in

concentration. Conversely, the mix may be limited to high scoring substituents, or their concentrations may merely be increased. The concentration changes need not be strictly proportional to the scores.

If a reference protein is very similar to the query protein, it should have a strong influence on the library composition, and, if its similarity is only modest, its influence should be weak. Likewise a drug lead which strongly resembles a candidate library component should strongly favor that library, and one which only weakly resembles it should imply a more modest enrichment.

One possible mathematical approach is shown below.

Let

$W_{\phi L}$ = original absolute weight of library type L in composite library [0..1]

S_p = similarity of query protein to reference protein [0..1]

S_d = similarity of reference protein drug d to library type L [0..1]

q_d = quality of drug d as drug lead in general [0..1]

then, for each L, the new relative weight w_L' adjusted for drug lead d of reference protein p is

$w'_L = (W_{\phi L} * (1-S_p)) + (W_{\phi L} * S_p * Q)$ where

$Q = (q_d * S_d) + (1-S_d)$.

If qd=1, then

$W'_L = (W_{\phi L} * (1-S_p)) + (W_{\phi L} * S_p * S_d)$

The new relative weights W'_L are converted to new absolute weights by dividing each by the sum of W'_L for all L.

Suppose that we were contemplating synthesis of a composite library composed of (a) benzodiazepines and (b) carbamates. In the absence of any guidance from the database, we might make the library 50% benzodiazepines and 50% carbamates. Hypothesizing that the target protein was 20% similar to a reference protein, and the reference protein had a known antagonist which was 90% similar to a benzodiazepine and 5% to a carbamate, we could recalculate weights as follows:

benzodiazepines

.5 x .8 = .4

.5 x .2 x .9 = .09
.4 + .09 = .49

carbamates
.5 x .8 = .4
.5 x .2 x .05 = .005
.4 + .005 = .405

.49 + .405 = .895
.49/.895 = .61 new benzodiazepine fraction
.405/.895 = .39 new carbamate fraction

Now assume that the target protein is 90% similar to the same reference protein. Then the calculation becomes

benzodiazepines
.5 x .2 = .1
.5 x .8 x .9 = .36
.1 + .36 = .46

carbamates
.5 x .2 = .1
.5 x .8 x .05 = .004
.1 + .004 = .104

.46 + .104 = .564
.46/.564 = .82 new benzodiazepine fraction
.104/.564 = .18 new carbamate fraction

If there is more than one relevant protein, we may first adjust the weights based on the most similar protein, then the new weights based on the next most similar protein, and so on. For example, let us assume that the most similar protein had a similarity of 80%, and an antagonist which was 90% benzodiazepine-type and 5% carbamate-type. If so, the adjusted weights would, as stated above, .82 benzodiazepine and .18 carbamate. If the next best protein were 20% similar, and its antagonist was 10% benzodiazepine-like and 70% carbamate-like, we would have

benzodiazepines

.82 x .8 = .66

.82 x .2 x .1 = .02

.66 + .02 = .68

carbamates

.18 x .8 = .14

.18 x .2 x .7 = .03

.14 + .03 = .17

.68 + .17 = .85

.68/.85 = .8 new benzodiazepine fraction

.17/.85 = .2 new carbamate fraction

If, for a given reference protein, more than one drug lead is known, the leads could be considered in descending order of structural similarity with a proposed library component, i.e., most similar lead first.

It will be appreciated that this is only one of many possible methods of adjusting library composition for reference protein/drug lead similarity.

Also, the user may decide to simply screen the most similar of the candidate simple combinatorial libraries. This is equivalent to giving it a weight of 1 and the others a weight of 0.

Drug leads are not equal in value. The drugs will vary in potency, side effects, deliverability, residence time, ease of synthesis, cost of production etc. Those factors which the chemist wishes to consider may be subsumed into a quality factor (q_d) ranging from 0 to 1, with unity being the most desirable sort of lead. A quality factor may be assigned by any rational method. If only potency is considered, then the most potent drug in the database could be assigned a value of 1, and the other drugs assigned relative values based on the logarithms of their potencies. Thus, if the best drug has an IC50 of 10^{-12}, a drug with an IC50 of 10^{-6} might have a q_d of 1/6. Of course, any function which converts potencies to a zero to one scale in which higher potencies yield higher values might be used.

By means of a quality factor, "negative teachings" may also

be taken into account. If the database includes compounds known not to inhibit a "retrieved" reference protein, then the library could be designed to reduce the representation of similar compounds. The formula given previously will do that, since w'_L is lower is Q is lower, and Q is lower if q_d is lower. The extent to which q_d influences w'_L is dependent on both S_p and on s_d. Other formulae could be used, e.g.

$w'_L = (w_{\phi L} * (1-sp) + (w_{\phi L} * Sp * S_d)^{Q'}$ where

$Q' = 1/Q$ and $0 < q_d <$ infinity;

(the neutral point is the $q_d = 1$).

In a similar manner, possible substituents for the library can be evaluated for similarity to the database-generated drug leads.

Examples of candidate simple libraries which might be evaluated include derivatives of the following:

Cyclic Compounds Containing One Hetero Atom

Heteronitrogen

pyrroles

pentasubstituted pyrroles

pyrrolidines

pyrrolines

prolines

indoles

beta-carbolines

pyridines

dihydropyridines

1,4-dihydropyridines

pyrido[2,3-d]pyrimidines

tetrahydro-3H-imidazo[4,5-c] pyridines

Isoquinolines

tetrahydroisoquinolines

quinolones

beta-lactams

azabicyclo[4.3.0]nonen-8-one amino acid

Heterooxygen

furans

tetrahydrofurans

2,5-disubstituted tetrahydrofurans

pyrans

hydroxypyranones
tetrahydroxypyranones
gamma-butyrolactones
Heterosulfur
sulfolenes
Cyclic Compounds with Two or More Hetero atoms
Multiple heteronitrogens
imidazoles
pyrazoles
piperazines
diketopiperazines
arylpiperazines
benzylpiperazines
benzodiazepines
1,4-benzodiazepine-2,5-diones
hydantoins
5-alkoxyhydantoins
dihydropyrimidines
1,3-disubstituted-5,6-dihydopyrimidine-2,4-diones
cyclic ureas
cyclic thioureas
quinazolines
chiral3-substituted-quinazoline-2,4-diones
triazotes
1,2,3-triazoles
purines
Heteronitrogen and Heterooxygen
dikelomorpholines
isoxazoles
isoxazolines
Heteronitrogen and Heterosulfur
thiazolidines
N-axylthiazolidines
dihydrothiazoles
2-methylene-2,3-dihydrothiazates
2-aminothiazoles
thiophenes
3-amino thiophenes

4-thiazolidinones
4-melathiazanones
benzisothiazolones

For details on synthesis of libraries, see Nefzi, et al., Chem. Rev., 97:449-72 (1997), and references cited therein.

In designing the library, one may consider not only the drug leads retrieved for the high similarity reference proteins, but also structures which are analogues of the BioKey peptides or nucleic acid aptamers known to bind the query protein and/or the higher-ranked reference proteins.

The present invention is useful, not only in designing a library, but also in deciding which of several query proteins to target. It may be desirable to inhibit a biochemical pathway. Several different proteins may be known to mediate that pathway. To decide which one to develop drugs for, each protein might be used to query the database. The one with the most potent and specific drug leads, or with the drug lead having the greatest resemblance to a readily available combinatorial library, becomes the target of choice.

Examples

Hypothetical Example 1 - Generation of DNA Aptamer

The protein is immobilized, e.g, on filter paper, and subjected to binding by a library of random oligonucleotides. The oligonucleotides may be DNA, RNA, or a DNA or RNA analogue, e.g., in which the sugar functionality of the nucleotide is modified at the 2′ position with substituents such as fluoro-, amino-, or methoxy-. The members of the library that bind to the target protein are separated and amplified using the polymerase chain reaction or other amplification scheme. The amplified pool is bound to the immobilized protein again, and the strong binding fraction is selected. This process is repeated if need be, e.g., 5 - 15 times, until ligands of the desired affinity are obtained. The affinity of the pool may be determined after each cycle. When the desired affinity is reached, the oligonucleotides are cloned into suitable host cells, e.g., bacteria, and the sequences are determined by automated methods.

Hypothetical Example 2 - Determination of contact sites of aptamer on target

The nucleic acid aptamers for a specific target are resynthesized and 3′- or 5′-labeled with ^{32}P. Each radiolabeled oligomer is then mixed with an excess of the target protein and 100 - 500 μM $Pt_2(pop)_4^{4-}$ and photolyzed at wavelengths between 350 and 500 nm until the reaction with $Pt_2(pop)_4^{4-}$ is complete. The radiolabeled oligomer is then precipitated. A parallel control reaction is run without the protein. The aptamers photolyzed with and without the protein are then loaded onto a polyacrylamide sequencing gel. The nucleotides where there is significant reaction without the protein but not with the protein are then classified as contact sites. This process is repeated for all of the aptamers for a given target and shows a high degree of similarity in the contact sites between aptamers.

Suppose a DNA library is screened for binding to lysozyme (cp. R. Diamond, J. Mol. Biol. 82:371 (1974)) and yields an aptamer with the sequence 5′- TAGCTGGCCAAAGTGCGAACACGGCCTTG.

The secondary structure is predicted by standard methods (described above), which show that the bold bases are paired to give a seven-base stem with an AAA bulge and a CGAA loop.

This sequence is then synthesized and radiolabeled on the 5' end with ^{32}P. The labeled oligomer is reacted with $Pt_2(pop)_4^{4-}$ by photolysis at 400 nm in phosphate buffered saline and analyzed on a sequencing gel, which shows that $Pt_2(pop)_4^{4-}$ induces scission of the oligonucleotide at every base in the sequence, giving a ladder of bands for the oligomer with approximately the same extent of scission at each nucleotide in the sequence.

The reaction is then repeated in the presence of enough protein to cause the majority of the oligonucleotide to be bound to the protein. The $Pt_2(pop)_4^{4-}$ exhibits some reaction with the protein, so the concentration of the $Pt_2(pop)_4^{4-}$ must be higher than in the reaction of the oligonucleotide alone, and the reaction must be performed in a short time so that alteration of the structure of the nucleoprotein complex due to damage of the protein by $Pt_2(pop)_4^{4-}$ is not a factor. Nevertheless, only the oligonucleotide is labeled, so only scission of the protein is detected. The scission pattern visualized on a sequencing gel then shows the same relative reactivity at the nucleotides that do not contact the protein as in the reaction of the oligomer alone, and greatly attenuated reactivity at sites protected by the protein. Studies on crystallographically characterized DNA-protein complexes show that the experiment faithfully indicates the sites of contact of the protein on DNA (see Breiner already cited).

The "footprint" of the protein on the DNA is determined by quantitating the extent of cleavage at each nucleotide in both reactions to form two histograms, normalizing the two histograms so that nucleotides that are clearly outside the binding site give the same intensity, and then assigning contact sites as those where addition of the protein attenuates the relative intensity.

For the hypothetical lysozyme aptamer described above, cleavage without the protein will give approximately the same intensity at all nucleotides. The cleavage intensity in the reaction with the protein is then normalized for the nucleotides

on either end, and the sites where less cleavage occurs is counted as a protein site. For example, if relatively less cleavage is observed at the underlined sites: 5'-TAGCTGGCCAAAGTGCGAACACGGCCTTG, then the cleavage for example on the end nucleotides and in the CGAA loop would be relatively the same with and without the protein, while cleavage in the stem and AAA bulge would be protected by the protein. This result would imply that the protein recognizes the AAA bulge and flanking stem, and this structure would be used to search for proteins that bind similar functionality sequences.

Hypothetical Example 3: Comparison of Two Aptamers

The RNA aptamers shown in Scheme 1A (below) were isolated by SELEX for the reverse transcriptase from Feline Immunodeficiency Virus (Chen, H.; McBroom, D.G.; Zhu, Y.-Q; Gold, L.; North, T.W. Biochemistry 1996, 35, 6923-6930) and from the ribosomal L22 protein associated with Epstein-Barr Virus (Dobbelstein, M.; Shenk, T. J. Virology 1995, 69, 8027-8034). Suppose the two resulting aptamer-protein complexes were subjected to footprinting with Pt-pop or some related method and that the underlined bases were deemed to make contacts with the protein, as described in the previous example. Within the contact regions, the bold bases are clearly involved in base pairing to make two hairpins with four-base stems. From this point, the two functionality sequences shown in B can be deduced according to the codes, with an At corresponding to an A contacting the protein that is base paired to a T on the opposite strand, and Ao corresponding to a single-stranded A, and so on. This particular example does not contain mismatches.

The homology score is determined by comparing the two sequences according to the (illustrative) scoring system shown in C. By this scheme, a single-stranded residue is given a maximum score of 2, a mismatch is given a maximum score of 1.8, and a base pair is given a maximum score of 1. The longer of the two sequences is then chosen as the "parent sequence", and the maximum score is calculated, which in this case is 26. This sequence is then aligned with the homolog to be compared, allowing for gaps if necessary; it is important to choose the longer sequence as the parent sequence because this imposes an

implicit penalty for gaps. The score is then computed at each nucleotide. For a perfect match, the score is the same as the maximum score. For a partial match, the score is 0.5 for transposition of a base pair (i.e., substitution of Ta for At), 0.5 for a base pair/mismatch pair where the same base contacts the protein, and 0.9 for a two mismatches where the contacted base is the same. The summed comparison score is then divided by the maximum total score. The comparison then of the two aptamers in Scheme 2 gives a score of 11/26 or 0.423. (The score would have been 4/26 for an identity matrix).

DNA aptamers would be scored in an identical manner except that T would replace U. Related strategies might involve changing in any of the maximum score numbers for single strands, mismatches, and base pairs or any of the partial match scores. It also need not be the case that all four nucleotides (A, T/U, G, C) have the same maximum score or partial match scores. Also, the individual scores could be summed instead as the squares of all the scored and then divided by the sum of the squares of the maximum score.

Scheme 1 - Aptamer Comparison

A

5'-CAAA**CUGGGUUACAUUUCCAG**UACAGCA - Dobbelstein 1995

5'-GUACCGAAUGUGCUUUUC**GGCC**GAUUUUUGGCCCCUGCAG - Chen 1996

B

Parent Sequence

Cg-Ua-Gc-Gc-Go-Uo-Uo-Ao-Co-Ao-Uo-Uo-Uo-Cg-Cg-Au-Gc

Homolog

Gc-Gc-Cg-Cg, Go-Ao-Uo-Uo-Uo-Uo-Uo-Gc-Gc-Cg-Cg = 11/26

.5 0 .5 .5 0 2 0 0 2 2 2 .5 .5 0 .5 = 11

1 1 1 1 2 2 2 2 2 2 2 2 2 1 1 1 1 = 26

C

Perfect Matches

Ao = 2 Uo = 2 Go = 2 Co = 2

Au = 2 Ut = 1 Gc = 1 Cg = 1

Aa = 1.8 Uu = 1.8 Ga = 1.8 Ca = 1.8
Ag = 1.8 Ug = 1.8 Gu = 1.8 Cu = 1.8
Ac = 1.8 Uc = 1.8 Gg = 1.8 Cc = 1.8

Partial Matches
Au/Ua = 0.5,
Au/An = 0.5 (n <> u)
An/An' = 0.9 (n <> n')
Note: Only the case of the contacted base being A is shown, but it could be T, G or C, too. Thus, Cg/Gc would be a partial match, with a value of 0.5.

Hypothetical Example 4 - Determination of Reactivity Descriptors

Suppose Ru(tpy)(bpy)O^{2+} has a rate constant with methionine of 15 M^{-1} s^{-1} and the rate constant with all of the other amino acids is negligible, and suppose that $Pt_2(pop)_4^{4-}$ has a Stern-Volmer rate constant for tryptophan that is 1.0×10^7 M^{-1} s^{-1} and that all of the other amino acids give negligible rate constants. Now glutamine synthetase and lysozyme are reacted with the two compounds in the folded, unfolded, and BioKey-bound states. The BioKey peptides are engineered to avoid the inclusion of methionine and tryptophan so that cross-reaction with the BioKey is not an issue. The rate constants are normalized to the number of the most reactive residue for each reagent, i.e., the Ru(tpy)(bpy)O^{2+} rate constant is in moles of methionine and the $Pt_2(pop)_4^{4-}$ rate constant is in moles of tryptophan. Now the rate constants are measured in the three states to give (hypothetically):

protein	state	K(Ru(tpy)(bpy)O^{2+})[a]	k($Pt_2(pop)_4^{4-}$)[b]
glutamine synthetase	unfolded	12 M^{-1} s^{-1}	$9x10^6$ M^{-1} s^{-1}
	folded	6 M^{-1} s^{-1}	$5x10^5$ M^{-1} s^{-1}
	BioKey-bound	1 M^{-1} s^{-1}	$5x10^5$ M^{-1} s^{-1}
lysozyme	unfolded	12 M^{-1} s^{-1}	$9x10^6$ M^{-1} s^{-1}
	folded	0.1 M^{-1} s^{-1}	$3x10^6$ M^{-1} s^{-1}
	BioKey-bound	0.1 M^{-1} s^{-1}	$1x10^4$ M^{-1} s^{-1}

[a]Given as M^{-1} s^{-1} where the molar concentration is per total methionine in the two proteins. [B]Given as M^{-1} s^{-1} where the molar concentration is per total tryptophan.

For the unfolded proteins, the rate constants are close to those for the amino acids themselves but slightly attenuated due to some steric constraints imposed by inclusion in the linear polymer. For glutamine synthetase, there are 16 methionines per subunit, and eight of these are surface-exposed and therefore oxidizable by $Ru(tpy)(bpy)O^{2+}$, the rate constant for oxidation of folded glutamine synthetase then drops by one-half. Levine et al. also found that eight methionines could be oxidized in the folded protein (Levin 1995). As found by Levine, the eight surface methionines ring the outside of the binding site, so most of these methionines would be blocked by the bound BioKey, so the rate constant drops (hypothetically) to one-sixth the value of the folded protein. For $Pt_2(pop)_4^{4-}$, oxidation of glutamine synthetase relies on reaction of the two tryptophans per subunit, both of which are at the interface where subunits bind and therefore protected in the folded protein. One of the two tryptophans is slightly exposed, so the rate constant drops (hypothetically) to about 6% of it's initial value. The slightly exposed tryptophan is not near the binding site, so binding the BioKey does not alter the rate constant.

For lysozyme, there are two methionine residues, both of which are buried in the protein. Therefore, the $Ru(tpy)(bpy)O^{2+}$ rate constant drops (hypothetically) to 1% of it's original value. The two residues are not near the binding site, so BioKey binding has no effect. There are six tryptophans that can react with $Pt_2(pop)_4^{4-}$ and two of these are exposed and in the binding site. Therefore, the folded rate constant is (hypothetically) one-third that of the unfolded protein, and BioKey binding dramatically attenuates the oxidation of the folded protein.

The reactivity descriptors are then calculated by comparing the difference in the folded and unfolded or the BioKey and folded rate constants normalized by the unfolded rate constant. Thus, a descriptors can be calculated as descriptor (folded) = (folded-unfolded)/unfolded and descriptor (BioKey) = (BioKey - unfolded)/unfolded. This linear formula will weight heavily residues that are not occluded very much by folding or BioKey binding and will obscure subtle differences in folding. Therefore, an advantageous descriptor formula would be to

calculate the difference in square roots, i.e.:

descriptor(folded) = [(rate folded)$^{1/2}$ - (rate unfolded)$^{1/2}$]/(rate unfolded)$^{1/2}$

descriptor(BioKey) = [(rate BioKey)$^{1/2}$ - (rate unfolded)$^{1/2}$]/(rate unfolded)$^{1/2}$

Descriptors calculated in this manner from the data shown above are:

protein	state	descriptor Ru	descriptor Pt
glutamine synthetase	folded	0.292	0.764
	BioKey	0.711	0.764
lysozyme	folded	0.908	0.422
	BioKey	0.908	0.967

Examining the descriptors shows that glutamine synthetase has a large amount of surface methionine and much of the surface methionine is near the active site, that glutamine synthetase has a some surface tryptophan not near the active site, that lysozyme has very little surface methionine, and that lysozyme has surface tryptophan in the active site.

This example is shown for the case where the two reagents exhibit a reactivity profile that is 100% for a single amino acid, i.e. 100% methionine for $Ru(tpy)(bpy)O^{2+}$ and 100% tryptophan for $Pt_2(pop)_4^{4-}$, and 0% for all other amino acids. In these cases, the concentration of protein used to calculate the rate constants is just figured by multiplying the concentration of the protein times the fraction of the reactive amino acid in the sequence. If a given reagent had a reactivity profile of 40% isoleucine, 40% leucine, and 20% alanine, the protein concentration would be determined as the total protein concentration times ((0.40 times the fraction of the total amino acids which are of isoleucine residues) plus (0.40 times the fraction which are leucine residues) plus (0.20 times the fraction which are alanine residues)).

Hypothetical Example 5 - Using Similarities to Improve Lead Generation

Suppose the database exists with both aptamer descriptors

and reactivity descriptors and includes the proteins listed in the two previous examples (reverse transcriptase from FIV, ribosomal L22 protein, glutamine synthetase, and lysozyme) and many other proteins. Then aptamers and reactivity descriptors are determined for a new protein (protein X). Protein X is compared to the other proteins in the database and is determined to be 60% similar to lysozyme, 20% similar to ribosomal L22 protein, 8% similar to glutamine synthetase and <3% similar to all of the other proteins in the database. (Note that these percentages may sum up to less than or to greater than 100%.) So the known inhibitors of the proteins with a >3% similarity to protein X are tabulated and used to predict leads for protein X. Suppose that lysozyme is inhibited by a quinazolinone and a benzodiazepine; ribosomal L22 protein is inhibited by a natural product; and glutamine synthetase is inhibited by a benzodiazepine. The benzodiazepine and quinazolinone compounds may be synthesized as libraries (see review in A. Nefzi, J.M. Ostresh, R.A. Houghten, Chem. Rev., 97:449 (1997)). It will probably be the most efficient to screen the entire quinazolinone and benzodiazepine libraries since these compounds are part of larger libraries already and since the selection of these compounds by the database predicts a general ability of the parent scaffolds to bind protein X. In contrast, the natural product 3 is likely part of a large library of discrete compounds that can be screened individually. In this case, structural descriptors can be calculated for 3 and compounds within a certain similarity radius can be selected for screening according to known compound descriptor methods, as in Cummins et al. 1996 or Matter 1997.

This invention covers not only the recited preferred embodiments, but also all possible combinations thereof. The recitation of a numerical range should be deemed a recitation also of all possible subranges. If a class is recited, this invention also extends to the individual members and subclasses and to all possible combinations of member or subclasses of that class.

Bibliography

1. Cohen, J. (1997). "The Genomics Gamble." Science **275**: 770.

2. Strauss, E. J. and S. Falkow (1997). "Microbial Pathogenesis: Genomics and Beyond." Science **276**: 707-711.

3. Ellman, J., B. Stoddard and J. Wells (1997). "Combinatorial Thinking in Chemistry and Biology." Proc. Natl. Acad. Sci. USA **94**: 2779-2782.

4. Plunkett, M. J. and J. A. Ellman (1995). "Solid-Phase Synthesis of Structurally Diverse 1,4-Benzodiazapine Derivatives Using the Stille Coupling Reaction." J. Am. Chem. Soc. **117**: 3306-3307.

5. Kauvar, L. M., D. L. Higgins, H. O. Villar, J. R. Sportsman, A. Engqvist-Goldstein, R. Bukar, K. E. Bauer, H. Dilley and D. M. Rocke (1995). "Predicting Ligand Binding to Proteins by Affinity Fingerprinting." Chem. Biol. **2**: 107-118.

6. Cummins, D. J., C. W. Andrews, J. A. Bentley and M. Cory (1996). "Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds." J. Chem. Inf. Comput. Sci. **36**: 750-763.

7. Patterson, D. E., R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger (1996). "Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors." J. Med. Chem. **39**: 3049-3059.

8. Matter, H. (1997). "Selecting Optimally Diverse Compounds from Structure Databases: A Validation Study of Two-Dimensional Molecular Descriptors." J. Med. Chem. **40**: 1219-1229.

9. Villar, H. O. and L. M. Kauvar (1994). "Amino Acid Preferences at Protein Binding Sites." FEBS Lett. **349**: 125-130.

10. Levine, R. L., L. Mosoni, B. S. Berlett and E. R. Stadtman (1996). "Methionine Residues as Endogenous Antioxidants in Proteins." Proc. Natl. Acad. Sci. USA **93**: 15036-15040.

11. Hoyer, D., H. Cho and P. G. Schultz (1990). "A New Strategy for Protein Cleavage." J. Am. Chem. Soc. **112**: 3249-3250.

12. Schepartz, A. and B. Cuenoud (1990). "Site-Specific Cleavage of the Protein Calmodulin Using a Trifluoperazine-Based Affinity Reagent." J. Am. Chem. Soc. **112**: 3247-3249.

13. Weinstein, J. N., T. G. Myers, P. M. O'Connor, S. H. Friend, A. J. Fornace, K. W. Kohn, T. Fojo, S. E. Bates, L. V. Rubinstein, N. L. Anderson, J. K. Buolamwini, W. W. van Osdol, A. P. Monks, D. A. Scudiero, E. A. Sausville, D. W. Zaharevitz, B. Bunow, W. N. Viswanadhan, G. S. Johnson, R. E. Wittes and K. D. Paull (1997). "An Information-Intensive Approach to the Molecular Pharmacology of Cancer." Science **275**: 343-349.

14. Kier, L. B. and L. H. Hall (1976). Molecular Connectivity and Drug Research. New York, Academic Press.

15. Klebe, G. and U. Abraham (1993). "On the Prediction of Binding Properties of Drug Molecules by Comparative Molecular Field Analysis." J. Med. Chem. **36**: 70-80.

16. Barton, J. K. (1994). Metal/Nucleic-Acid Interactions. Bioinorganic Chemistry. I. Bertini, H. B. Gray, S. J. Lippard and J. S. Valentine. Mill Valley, CA, University Science Books: 505-584.

17. Thorp, H. H. (1995). "Electron-, Energy-, and Atom-Transfer Reactions Between DNA and Metal Complexes." Adv. Inorg. Chem. **43**: 127-177.

18. Cheng, C.-C., J. G. Goll, G. A. Neyhart, T. W. Welch, P. Singh and H. H. Thorp (1995). "Relative Rates and Potentials of Competing Redox Processes During DNA Cleavage: Oxidation Mechanisms and Sequence-Specific Catalysis of the Self-Inactivation of Oxometal Oxidants." J. Am. Chem. Soc. **117**: 2970-2980.

19. Neyhart, G. A., C.-C. Cheng and H. H. Thorp (1995). "Kinetics and Mechanism of the Oxidation of Sugars and Nucleotides by Oxoruthenium(IV): Model Studies for Predicting Cleavage Patterns in Polymeric DNA and RNA." J. Am. Chem. Soc. **117**: 1463-1471.

20. Carter, P. J., C.-C. Cheng and H. H. Thorp (1996). "Oxidation of DNA Hairpins by Oxoruthenium(IV): Effects of Sterics and Secondary Structure." Inorg. Chem. **35**: 3348 - 3354.

21. Tullius, T. D. and B. A. Dombroski (1986). "Hydroxyl Radical "Footprinting": High-Resolution Information about DNA-Protein Contacts and Application to Lambda Repressor and Cro Protein." Proc. Natl. Acad. Sci. USA **83**: 5469-5473.

22. Kalsbeck, W. A., N. Grover and H. H. Thorp (1991). "Photolytic Cleavage of DNA by $Pt_2(pop)_4^{4-}$." Angew. Chem. Int. Ed. Engl. **30**: 1517.

23. Kalsbeck, W. A., D. M. Gingell, J. E. Malinsky and H. H. Thorp (1994). "Understanding the Interactions of [Pt2(pop)4]4- with Nucleic Acids: Photocatalytic Hydrogen Abstraction in Aqueous Solution." Inorg. Chem. **33**: 3313-3316.

24. Breiner, K. M., M. A. Daugherty, T. G. Oas and H. H. Thorp (1995). "An Anionic Diplatinum DNA Photocleavage Agent: Chemical Mechanism and Footprinting of Lambda Repressor." J. Am. Chem. Soc. **117**: 11673-11679.

25. Roundhill, M. D. (1985). "Excited-State Chemistry of Tetrakis(m-Pyrophosphito)diplatinum(II)." J. Am. Chem. Soc. **107**: 4354.

26. Kalyanasundaram, K. (1982). "Photophysics and Photochemistry of Diimine Complexes of Ruthenium(II)." Coord. Chem. Rev. **46**: 159.

27. Johnston, D. H., K. C. Glasgow and H. H. Thorp (1995). "Electrochemical Measurement of the Solvent Accessibility of Nucleobases Using Electron Transfer Between DNA and Metal Complexes." J. Am. Chem. Soc. **117**: 8933-8938.

28. Johnston, D. H. and H. H. Thorp (1996). "Cyclic Voltammetry Studies of Polynucleotide Binding and Oxidation by Metal Complexes: Homogeneous Electron-Transfer Kinetics." J. Phys. Chem. **100**: 13837-13843.

29. Johnston, D. H., T. W. Welch and H. H. Thorp (1996). "Electrochemically Activated DNA Oxidation." Metal Ions Biol. Syst. **33**: 297-324.

30. Scott, J. K. (1992). "Discovering Peptide Ligands with an Epitope Library." Trends Biochem. Sci. **17**: 241-245.

31. Kay, B. K., N. B. Adey, Y.-S. He, J. P. Manfredi, A. H. Mataragnon and D. M. Fowlkes (1993). "An M13 Library Displaying 38-amino-acid Peptides as a Source of Novel Sequences with Affinity to Selected Targets." Gene **128**: 59-65.

32. Kay, B. K. (1995). "Mapping Protein-Protein Interactions with Biologically Expressed Random Peptide Libraries." Persp. Drug Discovery Design **2**: 251-268.

33. Kay, B. K. and J. I. Paul (1995). "High-Throughput

Screening Strategies to Identify Protein-Protein Interactions." Mol. Div. 1: 139.

34. Stultz, L. K., R. A. Binstead, M. S. Reynolds and T. J. Meyer (1995). "Epoxidation of Olefins by [Ru(bpy)2(py)O]2+ in Acetonitrile Solution. A Global Analysis of the Epoxidation of trans-Stilbene." J. Am. Chem. Soc. 117: 2520-2532.

35. Marsh, E. N. (1995). "A Radical Approach to Enzyme Catalysis." Bioessays 17: 431-441.

36. Chao, C.-C., Y.-S. Ma and E. R. Stadtman (1997). "Modification of Protein Surface Hydrophobicity and Methionine Oxidation by Oxidative Systems." Proc. Natl. Acad. Sci. USA 94: 2969-2974.

37. Sinning, I., G. J. Kleywegt, S. W. Cowan, P. Reinemer, H. W. Dirr, R. Huber, G. L. Gilliland, R. N. Armstrong, X. Ji, P. G. Board, B. Olin, B. Mannervik and T. A. Jones (1993). "Structure Determination of Refinement of Human Alpha Class Glutathione Transferase A1-1, and a Comparison with the Mu and Pi Class Enzymes." J. Mol. Biol. 232: 192.

38. Napolitano, E. W., H. O. Villar, L. M. Kauvar, D. L. Higgins, D. Roberts, J. Mandac, S. K. Lee, A. Engqvist-Goldstein, R. Bukar, B. L. Calio, H. M. Jack and J. A. Tainer (1996). "Glubodies: Randomized Libraries of Glutathione Transferase Enzymes." Chem. Biol. 3: 359-367.

39. Pai, E. F., W. Kabsch, U. Krengel, K. C. Holmes, J. John and A. Wittinghofer (1989). "Structure of the Guanine-Nucleotide-Binding Domain of the Ha-ras Oncogene Product in the Triphosphate Conformation." Nature 341: 209.

40. Sicheri, F., I. Moarefi and J. Kuriyan (1997). "Crystal Structure of the Src Family Tyrosine Kinase Hck." Nature 385: 602.

41. Kussie, P. H., S. Gorina, V. Marechal, B. Elenbaas, J. Moreau, A. J. Levine and N. P. Pavletich (1996). "Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain." Science 274: 948.

42. Welch, T. W., G. A. Neyhart, J. G. Goll, S. A. Ciftan and H. H. Thorp (1993). "Thymidine-Specific Depyrimidination of DNA by Oxopolypyridylruthenium(IV) Complexes." J. Am. Chem. Soc. 115: 9311-9312.

43. Pratviel, G., J. Bernadou and B. Meunier (1995). "Carbon-

Hydrogen Bonds of DNA Sugar Units as Targets for Chemical Nucleases and Drugs." Angew. Chem. Int. Ed. Engl. **34**: 746-769.

44. Grover, N. and H. H. Thorp (1991). "Efficient Electrocatalytic and Stoichiometric Cleavage of DNA by Oxoruthenium(IV)." J. Am. Chem. Soc. **113**: 7030.

45. Macht, M., W. Fiedler, K. Kurzinger and M. Przybylski (1996). "Mass Spectrometric Mapping of Protein Epitope Structures of Myocardial Infarct Markers Myoglobin and Troponin T." Biochemistry **35**: 15633-15639.

46. Thompson, M. S. and T. J. Meyer (1982). "Kinetics and Mechanism of Oxidation of Aromatic Hydrocarbons by $Ru(tpy)(bpy)O^{2+}$." J. Am. Chem. Soc. **104**: 5070.

CLAIMS

1. A method of identifying drugs which mediate the biological activity of a protein of interest which comprises:

(a) determining the reactivity of the protein of interest, in one or more states, with one or more reagents, thereby obtaining a reactivity descriptor for said protein, and similarly determining comparable reactivity descriptors for one or more reference proteins, each reference protein having a biological activity known to be mediated by one or more known reference drugs;

(b) comparing the reactivity descriptor for the protein of interest with the reactivity descriptors of said reference proteins, identifying lead proteins, which are reference proteins whose reactivity descriptors are substantially similar to those of the protein of interest, and identifying lead drugs, which are reference drugs which mediate the biological activity of lead proteins,

(c) preparing a combinatorial compound library which is enriched for lead drugs and analogues thereof, and

(d) screening said combinatorial compound library for drugs which can mediate the biological activity of the protein of interest.

2. The method of claim 1 wherein the reactivity of the protein of interest, and of the reference proteins, is determined in both the folded and unfolded state.

3. The method of claim 1 wherein the reactivity of the protein of interest, and of the reference proteins, is determined in both the free state and in a ligand-bound state.

4. The method of claim 3 in which the ligand is an oligonucleotide.

5. The method of claim 3 in which the ligand is a peptide or a peptoid.

6. The method of claim 3 wherein the ligand is identified by screening a combinatorial library.

7. The method of claim 1 in which the protein of interest, and the reference proteins, are further characterized by the identification of one or more aptamers which bind said proteins, and the similarity of the protein of interest to each of said reference proteins is determined at least in part on the basis of the similarity of the respective aptamers which bind them.

8. The method of claim 7 in which the aptamers are olionunucleotides.

9. The method of claim 7 in which the protein of interest, and the reference proteins, are further characterized on the basis of the individual nucleotides of said aptamers which contact said proteins.

10. The method of claim 7 in which the protein of interest, and the reference proteins, are further characterized on the basis of the predicted or actual secondary structure of the aptamers which bind them.

11. A method of identifying drugs which mediate the biological activity of a protein of interest which comprises:

(a) determining the sequences of aptamers which bind the protein of interest and the sequences of aptamers which bind one or more reference proteins, each reference protein having a biological activity known to be mediated by one or more reference drugs, thereby obtaining aptamer descriptors for said proteins,

(b) comparing the aptamer descriptors for the protein of interest with the aptamer descriptor for each reference protein, identifying lead proteins, which are reference proteins whose

aptamer descriptors are substantially similar to those of the protein of interest, and identifying lead drugs, which are reference drugs which mediate the biological activity of lead proteins,

(c) preparing a combinatorial compound library which is enriched for lead drugs and analogues thereof, and

(d) screening said combinatorial compound library for drugs which can mediate the biological activity of the protein of interest.

12. The method of claim 11 in which the aptamers are olgonucleotides.

13. The method of claim 12 in which the protein of interest, and the reference proteins, are further characterized on the basis of the individual nucleotides of said aptamers which contact said proteins.

14. The method of claim 11 in which the protein of interest, and the reference proteins, are further characterized on the basis of the predicted or actual secondary structure of the aptamers which bind them.

FIG. 1

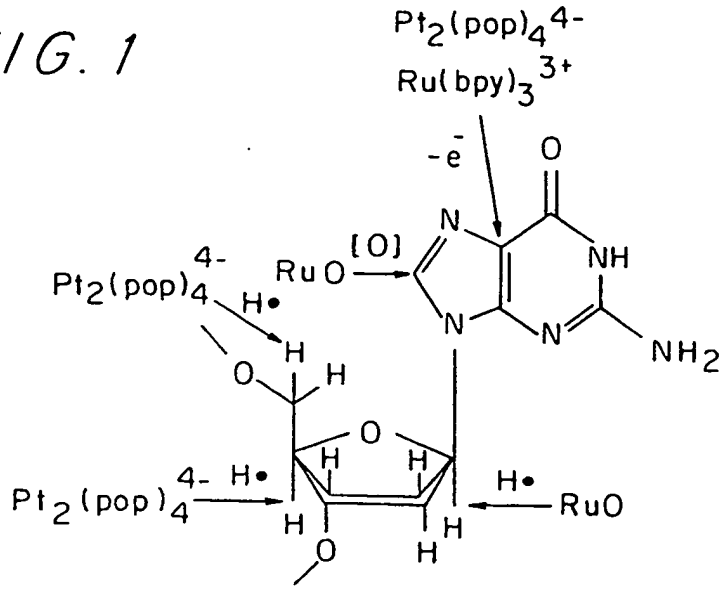
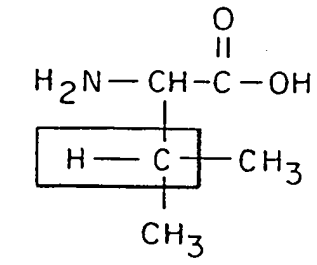
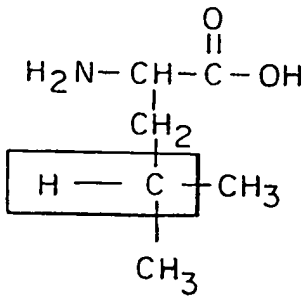


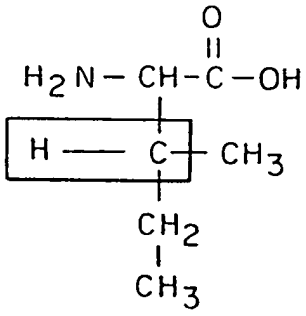
FIG. 2



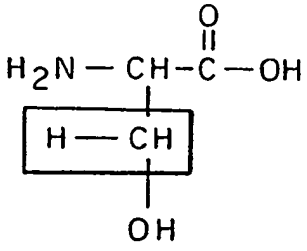
VALINE



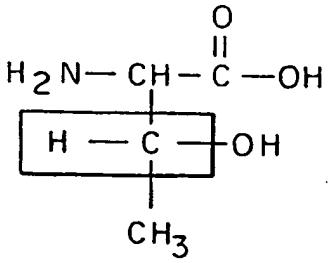
LEUCINE



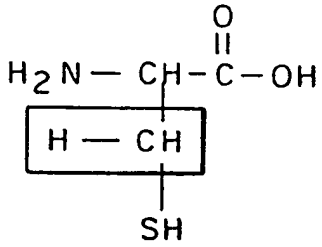
ISOLEUCINE



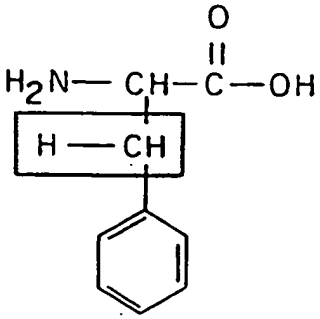
SERINE



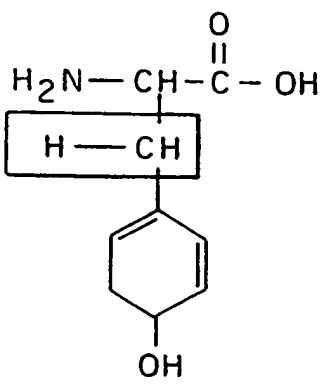
THREONINE



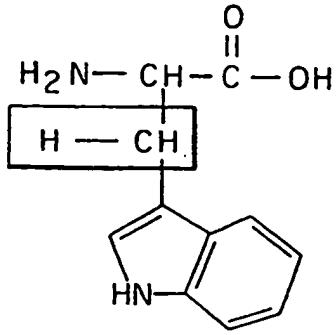
CYSTEINE



PHENYLALANINE



TYROSINE



TRYPTOPHAN

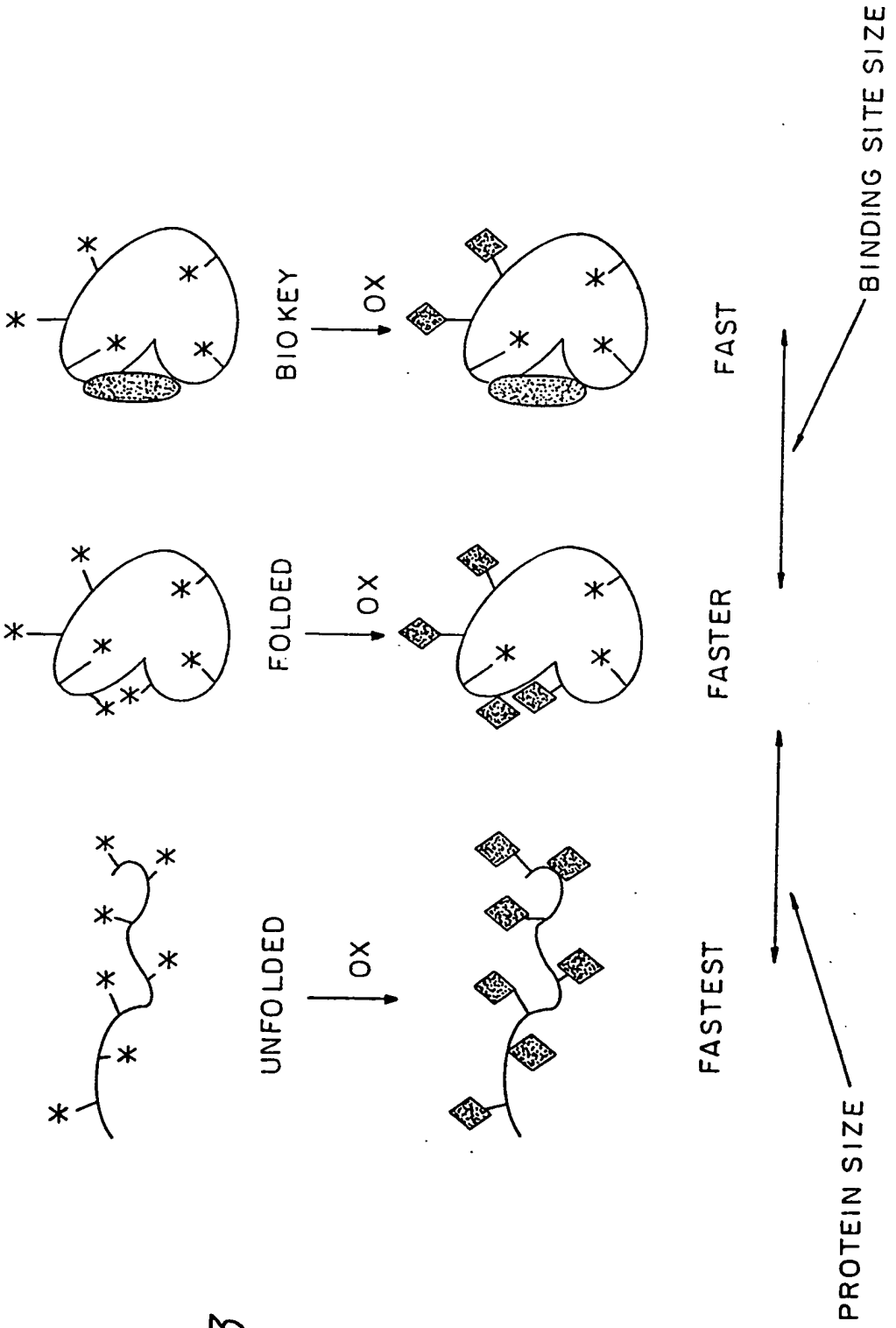

BIO KEY
OX
FOLDED
OX
UNFOLDED
OX
FAST
FASTER
FASTEST
BINDING SITE SIZE
PROTEIN SIZE

FIG. 3

FIG. 4

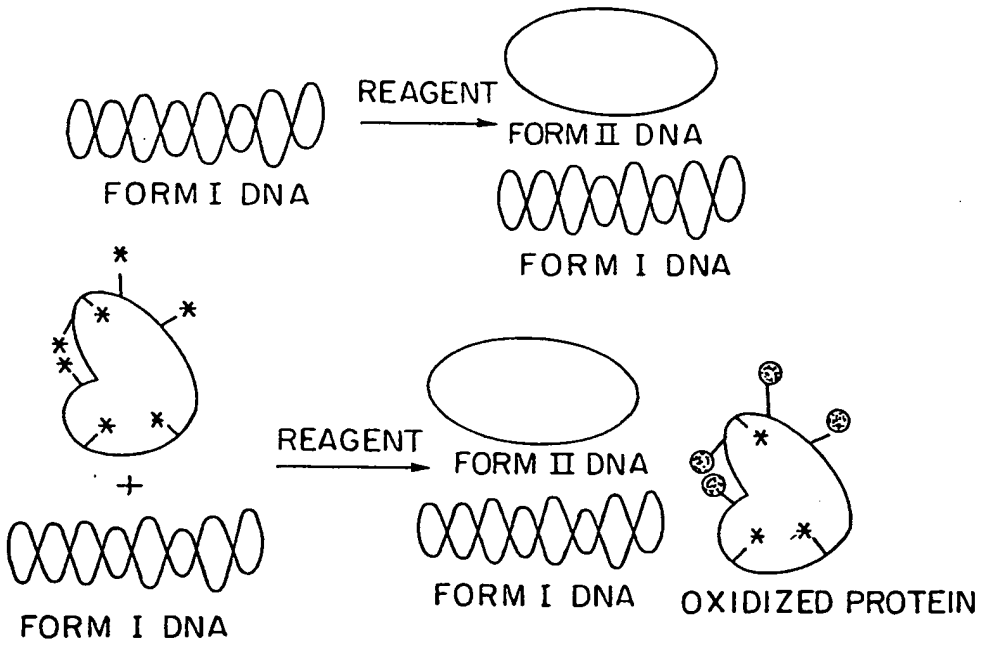
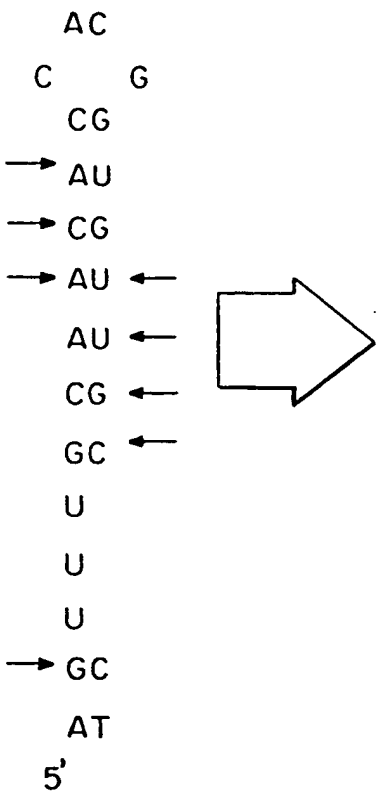

REAGENT
FORM II DNA
FORM I DNA
FORM I DNA
REAGENT
FORM II DNA
FORM I DNA
FORM I DNA
OXIDIZED PROTEIN

FIG. 5



	No	No	
No			No
	Nn-	Nn	
	Au-	Nn	
	Cg-	Nn	
	Au-	Ua	
	Nn-	Ua	
	Nn-	Gc	
	Nn-	Gc	
No			
No			
No			
	Gc-	Nn	
	Nn-	Nn	

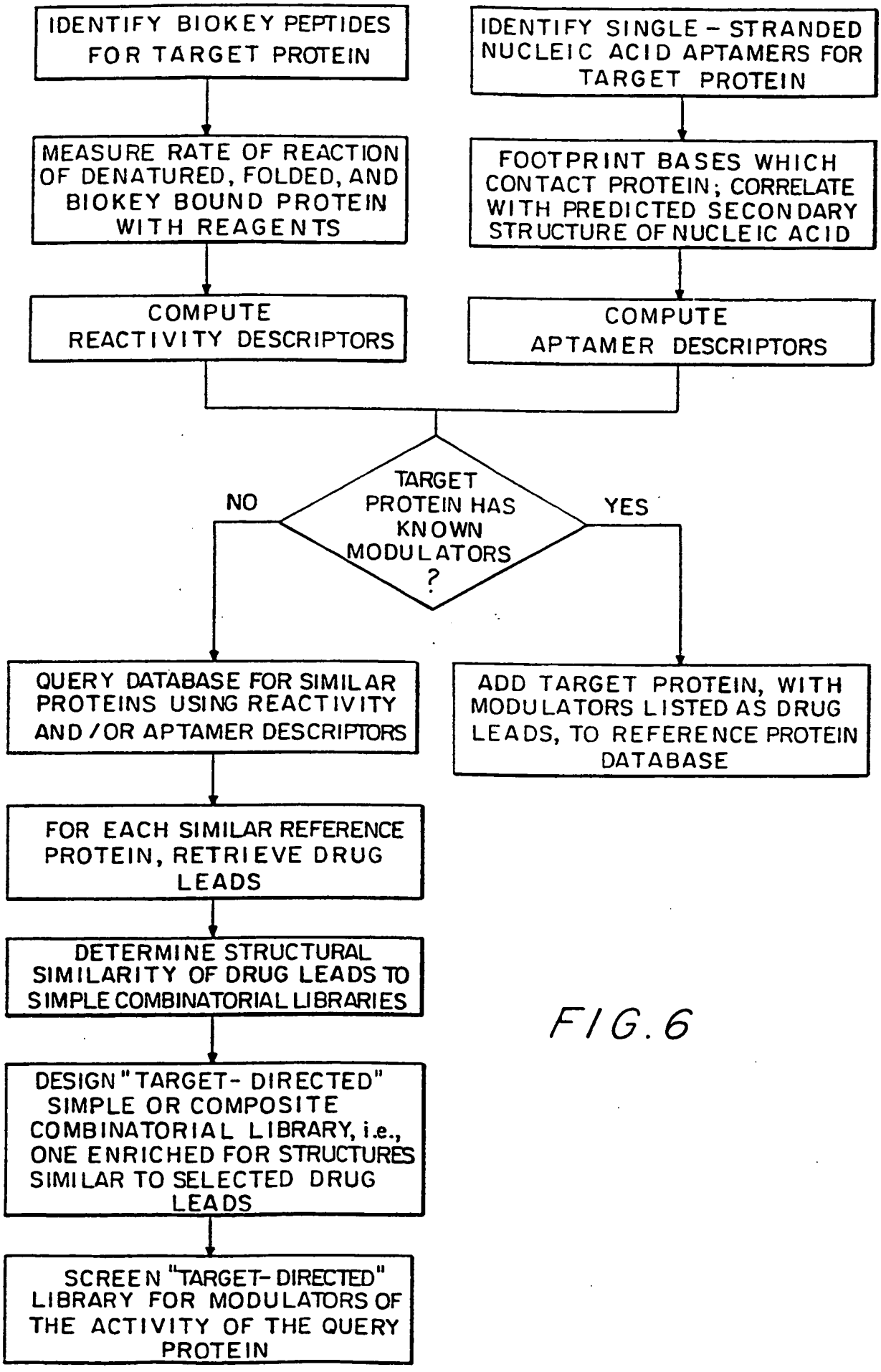

IDENTIFY BIOKEY PEPTIDES FOR TARGET PROTEIN
IDENTIFY SINGLE-STRANDED NUCLEIC ACID APTAMERS FOR TARGET PROTEIN
MEASURE RATE OF REACTION OF DENATURED, FOLDED, AND BIOKEY BOUND PROTEIN WITH REAGENTS
FOOTPRINT BASES WHICH CONTACT PROTEIN; CORRELATE WITH PREDICTED SECONDARY STRUCTURE OF NUCLEIC ACID
COMPUTE REACTIVITY DESCRIPTORS
COMPUTE APTAMER DESCRIPTORS
TARGET PROTEIN HAS KNOWN MODULATORS ?
NO
YES
QUERY DATABASE FOR SIMILAR PROTEINS USING REACTIVITY AND/OR APTAMER DESCRIPTORS
ADD TARGET PROTEIN, WITH MODULATORS LISTED AS DRUG LEADS, TO REFERENCE PROTEIN DATABASE
FOR EACH SIMILAR REFERENCE PROTEIN, RETRIEVE DRUG LEADS
DETERMINE STRUCTURAL SIMILARITY OF DRUG LEADS TO SIMPLE COMBINATORIAL LIBRARIES
DESIGN "TARGET-DIRECTED" SIMPLE OR COMPOSITE COMBINATORIAL LIBRARY, i.e., ONE ENRICHED FOR STRUCTURES SIMILAR TO SELECTED DRUG LEADS
SCREEN "TARGET-DIRECTED" LIBRARY FOR MODULATORS OF THE ACTIVITY OF THE QUERY PROTEIN

FIG. 6

INTERNATIONAL SEARCH REPORT

Interr nal Application No
PCT/US 98/15943

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G01N33/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,X	WO 97 42500 A (DIMENSIONAL PHARMACEUTICALS IN) 13 November 1997 see claims 1,3,16,30,74-90	1-14
Y	WO 95 32425 A (SMITHKLINE BEECHAM CORP ;YAMASHITA DENNIS SHINJI (US); WEINSTOCK J) 30 November 1995 see the whole document	1-14
Y	US 5 338 659 A (KAUVAR LAWRENCE M ET AL) 16 August 1994 see the whole document	1-14
Y	WO 93 01484 A (UNIV CALIFORNIA) 21 January 1993 see the whole document	1-14

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 December 1998	29/12/1998

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Routledge, B

1

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PCT/US 98/15943

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P,A	ELOFSSON M ET AL: "Tightening the nuts and bolts" TRENDS IN BIOTECHNOLOGY, vol. 16, no. 4, April 1998, page 147-149 XP004112297 see the whole document	1-14

INTERNATIONAL SEARCH REPORT

...information on patent family members

Inter... ...nal Application No
PCT/US 98/15943

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9742500 A	13-11-1997	AU 3205097 A	26-11-1997
WO 9532425 A	30-11-1995	EP 0763202 A	19-03-1997
		JP 10500951 T	27-01-1998
US 5338659 A	16-08-1994	AU 662001 B	17-08-1995
		AU 1987892 A	02-11-1992
		CA 2107474 A	03-10-1992
		EP 0581881 A	09-02-1994
		JP 6509865 T	02-11-1994
		WO 9217784 A	15-10-1992
		US 5674688 A	07-10-1997
		US 5679643 A	21-10-1997
		US 5763570 A	09-06-1998
WO 9301484 A	21-01-1993	AU 2408292 A	11-02-1993
		US 5436850 A	25-07-1995